Uncovering the cis-regulatory program of early human B-cell commitment and its implications in the pathogenesis of B-cell acute lymphoblastic leukemia

Authors: Núria Planell[1,2], Xabier Martínez-de-Morentin[1], Daniel Mouzo[1], David Lara-Astiaso[3], Amaia Vilas-Zornoza[3], Patxi San Martín-Uriz[3], Diego Alignani[4], Bruno Paiva[5], Alberto Maillo[6], Aleksandra Kurowska[6], Arantxa Urdangarin[1], Peri Noori[7], Asier Ortega-Legarreta[1], Mikel Hernaez[2], Vincenzo Lagani[6], Narsis Kiani[8,9], José I. Martín-Subero[10,11], Ricardo N. Ramírez[12], Jesper Tegner[13,14], Felipe Prosper[3,15], David Gomez-Cabrero[1,6]

[1]Translational Bioinformatics Unit, Navarrabiomed, Universidad Pública de Navarra (UPNA), Instituto de Investigación Sanitaria de Navarra (IdiSNA), Pamplona, Spain.
[2]Universidad de Navarra, Centro de Investigación Médica Aplicada (CIMA), Computational Biology Program, Instituto de Investigación Sanitaria de Navarra (IdiSNA), Pamplona, Spain.
[3]Hematology-Oncology Program, CIMA, Cancer Center Clínica Universidad de Navarra (CCUN), IdISNA, Pamplona, Spain. Centro de Investigación Biomédica en Red de Cáncer, CIBERONC.
[4]Flow Cytometry Core Facility, Centro de Investigación Médica Aplicada, Universidad de Navarra, CCUN, IDISNA, CIBERONC, Pamplona, Spain.
[5]Clínica Universidad de Navarra, Centro de Investigación Médica Aplicada, Universidad de Navarra, CCUN, IDISNA, CIBERONC, Pamplona, Spain.
[6]Bioscience Program, Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology KAUST, Thuwal 23955, Saudi Arabia.
[7]The Computational Medicine Group, Karolinska Institutet, Karolinska University Hospital Solna, Stockholm, Sweden
[8]Department of Oncology-Pathology, Center of Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.
[9]Algorithmic Dynamics Lab, Center of Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.
[10]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona,
[11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
[12]Department of Immunology, Harvard Medical School, Boston, MA 02115, USA.
[13]Computer, Electrical, and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology KAUST, Thuwal 23955, Saudi Arabia.
[14]Bioengineering Program, Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology KAUST, Thuwal 23955, Saudi Arabia.
[15]Hematology and Cell Therapy Service, Cancer Center Clínica Universidad de Navarra (CCUN), IdISNA, Pamplona, Spain

ABSTRACT

The early stages of the B-cell system are key for cellular immunity development, and alterations may lead to various disorders. Understanding the gene regulatory network (GRN) of this system is essential for studying healthy development and malignant transformations. To this end, we generated matched human data for chromatin accessibility and transcriptome in eight B-cell precursors, providing the first deep characterization of early B-cell differentiation, including the regulatory elements definition and the reconstruction of the GRN governing this process. Our data revealed ELK3 as a critical transcription factor (TF) in pro-B cells and uncovered their upstream regulators. We also identified MLXIP within the EBF1 regulators. Interestingly, modifications of enhancers preceding transcriptional changes were shown. Importantly, this resource helped uncover B-cell acute lymphoblastic leukemia (B-ALL) triggers, identifying pro-B and pre-B cells as inflection points of malignant transformation for some subgroups. The resource also explored the overlap of B-ALL susceptibility loci in the epigenomic profile. Overall, our study provides the most comprehensive atlas of early human B cell regulation (B-rex) at https://brex.shinyapps.io/brex/, a resource for understanding B cell differentiation in health and disease.

INTRODUCTION

The development of functional B cells requires appropriate regulation of dynamic transcription-factor networks. These regulatory networks are responsible for lineage-specific gene expression activation and lineage commitment. A complete description of the wiring, the logic, and the dynamics of the gene regulatory network (GRN) of B cells, is essential for understanding their development and how alterations in this process can lead to malignant transformations and other B-cell-related diseases[1–3]. Significant progress has been made in identifying transcription factors (TFs) driving the B cell differentiation[3]. PAX5, EBF1, IKAROS, PU.1, and E2A (TCF3) are key regulators widely studied[3–8]. Yet, understanding the gene regulatory mechanisms of this process remains incomplete.

Integrated analysis of gene expression and chromatin accessibility provided initial insights into cis-regulation of the mouse immune system[9]. A similar approach revealed regulatory programs in mixed-phenotype acute leukemias[10]. Other approaches studying the regulatory processes associated with the dynamic remodeling of the chromatin have been performed with promoter-capture Hi-C across human primary blood cell subpopulations, revealing high-resolution maps of promoter interactions[11]. Hi-C analysis across B-cell differentiation demonstrated global modulation of chromatin structures[12]. Valuable efforts mapping DNaseI single base-pair resolution footprints have provided high-resolution insights into the TF occupancy specific to tissues and cell types[13], contributing to the discovery of TF binding motifs in cis-regulatory elements (CREs).

Altogether, technological advances and current knowledge have enabled moving forward into a holistic view of gene regulation and eased the translation to pathological conditions. However, data on early B cells is primarily derived from studies in mice[9,14,15], and it is unclear whether the identified mechanisms are also valid in humans[3]. Moreover, deep profiling of early stages of human B cell differentiation in sufficient *depth* (detailed profiling) and *breadth* (enough number of individuals) is lacking, and current single-cell experiments did not yet provide it.

Therefore, to decipher the GRN of early B cell differentiation in humans, we generated comprehensive profiling of the stepwise progression of B-cell precursors, including eight cell subpopulations (HSCs, CLPs, pro-B cells, pre-B cells, Immature B cells, and Transitional B cells as well as CD5+ and CD5- Naive B cells), from 13 healthy donors. Paired high-throughput data of chromatin accessibility and gene expression were generated. The combined analysis of these omics data allowed us to infer the regulatory elements of B cell differentiation and the resulting GRNs. The dynamic activity of regulatory elements was deciphered thanks to the granularity of our data. Based on the in-deep characterization of early B cell differentiation, we were able to reveal putative triggers of B-cell acute lymphoblastic leukemia.

RESULTS

We generated paired transcriptomic (RNA-Seq) and chromatin accessibility (ATAC-Seq) profiles from 8 cell subpopulations covering a differentiation window from HSCs to naïve B cells. Cell populations were isolated from healthy adult human bone marrow precursors and peripheral blood from the same individuals (n=13) following a predefined sorting strategy (**Figure S1** & **Table S1**). After data generation and processing, 79 high-quality RNA-Seq and 78 ATAC-Seq profiles, 59 of them paired, were retained (**Figure 1A, Table S2**).

Our data revealed 23,454 expressed genes (coding and non-coding) and 105,359 open chromatin regions (OCRs) (see Methods), giving the complete profiling of chromatin accessibility and transcriptional profile of early B cell lineage (**Figure 1B**, https://genome.ucsc.edu/s/TransBioLab/bcell_hg38). Moving forward into a mechanistic scenario, we aimed to decipher the gene regulatory elements derived from the RNA-Seq and ATAC-Seq coordination. We found 14,589 OCRs mapping in promoter regions, 75,774 in intronic or distal intergenic regions, and 15,026 in other regions (exon, 3'UTR, 5'UTR, etc.) (**Figure S2A**).

Looking for cis-regulatory elements (CREs), we focused on those OCRs mapping to a promoter, intronic or distal intergenic region. To establish the link between CREs and the gene(s) they regulate, we followed two approaches (**Figure S2B**): 1) For promoter OCRs, we directly annotated OCRs to genes based on TSS distance (<1kb), and 2) For intronic and distal intergenic regions (considered here as possible enhancer regions), we linked OCR to genes following the hypothesis that correlation across cell subpopulations between chromatin accessibility of an enhancer and the expression of a given gene denotes a likely functional connection[9] (see Methods). Our data granularity facilitated this parse and revealed 50,312 enhancer OCR-gene associations within a 1Mb window from TSS. Altogether, we identified 12,247 promoter and 22,123 enhancer regions regulating 12,805 genes (58% of genes showing only promoter regulation, 12% only enhancer regulation, and 30% both) (**Figure S2C-F**). For example, exploring identified candidate CREs (cCREs) for the PAX5 gene, a primary determinant of B cell commitment, we observed coordination between the accessibility of the promoter and the gene expression, showing a clear shift in pro-B cells for both omic layers (**Figure 1C, D**). Assuming this ATAC-RNA coordination at the enhancer level, we identified 46 candidate enhancer elements based on correlation for PAX5 (**Figure 1E-G**). Within the enhancer OCRs identified, two were located in intron 5, previously reported as potent enhancer region[16]. Among the rest, several matched with known interaction regions described in promoter capture Hi-C data from Naïve B cells[11]. Still, others were new (**Figure 1E, G** & **Table 3**), mainly those where the accessibility in Naive B cells was low (**Figure 1H** & **Figure S2I, J**). It is relevant to note that we observed a large overlap between epigenetic signatures (histone marks and genome compartments) and our OCRs. The overlap with H3K4me3, H3K4me1, and H3K27ac marks support the functionality of our OCRs as cCREs according to ENCODE classification[17] (**Figure 1E** & **Figure S2G, H**). Moreover, significant enrichment for open chromatin (A-type compartment from Hi-C) and signatures associated with active promoter-like elements (H3K4me3 and H3K27ac) and enhancer-like elements (H3K4me1 and H3K27ac) within identified cCREs was shown (**Figure S2G, H**)[12,17–19].

Next, aiming to characterize the functional role of chromatin accessibility through transcription factor (TF) activity, we mapped the TFs binding motifs in each of the 105,359 OCRs using the Vierstra database[13], resulting in 275 putative TFs (see Methods). Looking at the TFs associated with the cCREs of PAX5, we found well-known TFs, such as EBF1, FOXO1, TCF3, PAX5, IRF4/8[2,20], but also POU2F2 (OCT2), and KLF13, as new candidate trans-regulatory elements of PAX5.

Overall, the generated data allowed us to identify the whole cis- and trans-regulatory elements, giving over a million possible TF-OCR-gene interactions. This complete portrait of regulatory elements, publicly available in a user-friendly web interface for browsing (B-rex; https://brex.shinyapps.io/brex/), represents a rich playground to unravel gene regulatory networks (GRNs) associated with B cell differentiation.

Recovering the gene regulatory network driving human early B-cell differentiation

To uncover the GRN specifically associated with the B cell differentiation process, we first identified the genes and chromatin regions that were significantly modulated over cell transitions. As a result, we identified 11,080 differentially expressed genes (DEGs), capturing the key biological processes associated with B cell differentiation (**Figure S3D, E**) and 77,559 differential accessible regions (DARs) (**Figure S3F**). DARs were mainly associated with enhancer regions (intronic or distal intergenic), following the concept that chromatin accessibility, and more specifically enhancer elements, yield sharper distinctions between cell subpopulations than promoter regions or expression profiles (**Figure S3A, B**)[9,21].

Then, considering the identified DEGs and DARs responsible for B cell differentiation, we zoomed in on the regulatory elements of these features. From the whole set of regulatory elements previously identified, we kept the TF-OCR-gene interactions derived from the DEGs, their promoter regions, and the associated enhancer elements with differential accessibility, resulting in 1,151,024 TF-OCR-gene interactions. To reduce spurious and meaningless interactions, we excluded those where: TF shows no differential expression, no significant TF-gene correlation was observed, or meaningless interaction regarding TF-OCR-gene coordination appears (see Methods). As a result, 377,268 TF-OCR-gene interactions were retained to reconstruct the GRN of the B cell differentiation process, involving 169 TFs, 7,310 genes, and 16,074 OCRs (**Figure 2A-C**).

The inferred GRN included genes and OCRs able to draw the entire regulatory dynamics over differentiation behaving in accordance with established knowledge (**Figure 2A-B**). Genes and TFs related to homeostasis and stemness, such as HOXB6, MEIS1, CD34, TAL1, and CDH2, and antiapoptotic genes such as BCL2, were highly expressed in HSC, CLP, and pro-B cells. Then, in pro-B cells, began the up-regulation of several surface markers genes and TFs involved in the commitment of B cell lineage, including CD19, CD24, CD79A/B, PAX5, TCF3, EBF1, FOXO1, among others[3]. The NOTCH1 gene, which is linked to T cell commitment, was down-regulated in pre-B cells[22]. In contrast, we found a positive correlation between TF expression and activity (computed with the TF binding site (TFBS) activity score) (**Figure 2C** & **Figure S3G**).

Once the GRN governing B cell differentiation was defined, we explored the regulation dynamics over differentiation. To identify the key regulatory elements at each cell state, we generated a snapshot of the GRN for each cell subpopulation based on the gene expression and chromatin accessibility (see Methods). As a result, we obtained a specific GRN for each cell subpopulation. When all GRNs were taken together, they were able to recapitulate the entire B-cell differentiation trajectory (**Figure 2D & Figure S5A**). Next, we investigated the network topology, and to this end, we computed the degree and betweenness of the GRNs, which unraveled the node hub architectures (**Figure 2E**). The hub architectures mirrored the transcriptional activity of each TF, pinpointing the cell subpopulation where they have more impact. Focusing on nodes with high degree and betweenness values, we recovered the primary TFs involved in B cell commitment, differentiation, maintenance, and class-switch recombination: TCF3, BCL11A, IRF8, PAX5, REL, IRF4, EBF1, POU2F2, FOXO1, and LEF1[14,22–24]. Besides the well-known TFs, ELK3 was revealed as a critical TF in pro-B cells (**Figure 2F-G**). This ETS-factor, recently defined as a novel oncogene, is involved in angiogenesis, cell adhesion, migration, and invasion[25] and has been reported to increase in ETV6-RUNX1+ acute lymphoblastic leukemia (ALL) cases[24,26] and other carcinogenic processes[27–29]. Moreover, ELK3 has been associated with pro-B cell expansion[30]. Here, exploring the pathways enriched for those genes regulated by ELK3 on pro-B and pre-B cells (n=2,606 genes), previously described pathways, such as cell migration, adhesion, and angiogenesis, appeared significantly reduced in pre-B cells (**Figure S4A, B**), reinforcing the idea that ELK3 deregulation can play a role in B-cell malignancies and its inhibition may be an interesting therapy[25].

Altogether, we recovered a complex GRN of the entire differentiation process. Furthermore, according to gene expression and chromatin accessibility dynamics, a set of GRNs for each cell subpopulation was retrieved, revealing the regulatory elements governing each cell state.

Upstream regulation dynamics show the TF activation along the B cell differentiation.

After exploring the overall GRN dynamics globally, in-deep exploration of each network's genes was done under the downstream or upstream regulation approach. Over the whole GRN, a median of 25 TFs (IQR: 10-48) were identified as upstream regulators per gene, and a median of 957 downstream genes (IQR: 435-2,091) was associated per TF (**Figure S5B**). To explore the

potential of our GRN, we first investigated the upstream regulation of EBF1, a key determinant of B lineage specification and differentiation. Eighty-five putative TFs were revealed as responsible for EBF1 (**Figure 3A, C** & **Figure S5C**). The dynamics of GRN over the different cell subpopulations and the edge's wideness highlighted those TFs being more critical in regulating genes. This approach captured the well-known TFs controlling EBF1, including PAX5, TCF3, FOXO1, SPIB (Spi-1/PU.1 Related), RUNX1 and CEBPA[20], but revealed new candidate regulators such as LEF1, MEF2A, IRF4, TCF4, POU4F1, FOXO6 and MLXIP. Some have already been described as important players in B cell development[31] but not as EBF1 regulatory elements. Interestingly, MLXIP (MondoA), an MYC transcription factor interactome member associated with the proliferation control in B-ALL[32] but not previously related to B cell development, arose as a potential regulator. We identified MLXIP as highly expressed with increased TF activity in pro-B and pre-B cells (**Figure 3E**).

Next, we explored the upstream regulation of ELK3, a TF previously identified within the top TFs governing the B cell differentiation. Fifty putative TFs responsible for ELK3 regulation were revealed (**Figure 3B, D** & **Figure S5D**). Again, the GRN dynamic over cell subpopulations and the edge's wideness highlights key TFs, such as ERG, STAT5A/B, E2F3, ETS2, and ETV6, as activators and SPIB (Spi-1/PU.1 Related) and POU2F2 as inhibitors. Interestingly, ETV6 arose as a candidate regulator, which can shed light on the mechanistic understanding of the elevated activity of ELK3 described in ETV6-RUNX1-positive ALL[24]. An overview of the whole TFs upstream regulation is provided in **Figure S6**.

The deep exploration of the upstream regulation dynamic of B cell differentiation GRN painted a rich portrait that not only captured the known regulators but also identified new putative regulators and revealed their behaviour over the differentiation process.

Gene regulatory elements point to the putative cell origin and shed light on triggers of B-cell acute lymphoblastic leukemia.

Aiming to use the novel data generated in a pathologic context, we hypothesized that the deep characterization of the B-cell differentiation process will provide insights into the dysregulation of precursor B-cell differentiation-related diseases. As a case study, we investigated B cell precursor acute lymphoblastic leukemia (BCP ALL), the most common and highly heterogeneous childhood cancer[33]. Recently, the transcriptional landscape of BCP ALL has been defined under an international study of 1,223 individuals[34]. The different transcriptional subgroups of BCP ALL identified were associated with specific fusions or mutations of genes such as MEF2D fusions, TCF3–PBX1 fusions, and PAX5 mutations. Exploring the enrichment of the transcriptional profile of eight BCP ALL subgroups in our transcriptional signatures of B cell precursors, we could estimate the putative inflection point where the malignant transformation of B cell may have occurred (**Figure 4A, B**). According to data reported by Mehtonen et al.[24], we found ETV6-RUNX1 fusions significantly enriched in pro-B cells. Moreover, we found a significant association of MEF2D fusions in pre-B cells, TCF3-PBX1 mutations in pro-B, pre-B, and Immature B cells, DUX4 fusions in HSCs, and KMT2A fusions in CLPs. Fine-tuning this approach, we focused on the BCP ALL subgroups where the mutation involves a TF and explored the enrichment of the core genes derived from the TF's downstream regulation. This second approach pointed to the same inflection points and evidence of a strong association of TCF3-PBX1 fusions in pre-B cells (**Figure 4C**). Next, we explored 711 binding sites defined as the highest affinity binding sites for ETV6-RUNX1 fusion[35]. Again, an enrichment in pro-B cells was observed (**Figure 4D**). Going further in exploring ETV6-RUNX1 cases, the ELK3 gene, revealed by our GRN as a critical TF in pro-B cells (**Figure 2F, G**), showed increased activity in these patients[24] (**Figure 4E**). Moreover, the susceptibility locus rs17481869, associated with ETV6-RUNX1 patients[36], fell within a small OCR

specific to pro-B cells (**Figure 4F)**. Altogether, these data provided different sources of evidence pointing to the association of ETV6-RUNX1 with pro-B cells.

Based on that observation, we investigated additional ALL-associated SNPs. Our data revealed that the ALL-associated rs4762284 locus, localized in the first intron of ELK3[26], fell in a closed chromatin region, indicating no functional effect in the healthy context of B cell lymphopoiesis. As SNPs identified from GWAS studies may not represent themselves the functional SNP, but point to a region of the genome in linkage disequilibrium[37], we explored their proxies. Thus, we could identify rs2268508 and rs17736737 in an open region of chromatin (**Figure 4G**), which can shed light on the implications of these SNPs with the disease. Furthermore, from the list of SNPs associated with BCP ALL-related traits from the GWAS catalog (NHGRI-EBI), we identified 11 SNPs overlapping some of our OCRs (**Table S4**). From these, two (rs2239633 and rs35837782) fell within an OCR defined as cCRE. The rs2239633 polymorphism fell in the promoter of the CEBPE gene. Although controversy exists regarding its association with ALL[38], we identified the promoter region accessible throughout B cell differentiation (**Figure 4H**). and an increased expression of the CEBPE gene in ETV6-RUNX1 and hyperdiploid BCP ALL cases (**Figure 4I**). The rs35837782 polymorphism fell in an intronic OCR of the LHPP gene (**Figure 4J**). Our data revealed this OCR as a cCRE of LHPP, CHST15, and FAM53B-AS1 genes associated with MEF2A, TCF3, and SERBF1 TF motifs. To our knowledge, no data relates these genes with ALL or B cell differentiation; however, increased expression of these was shown in pro-B and pre-B cells and different subtypes of BCP ALL (**Figure 4K, L**).

Overall, these new insights on the BCP ALL illustrated the power of the data generated to obtain a deeper understanding of the molecular mechanisms underlying B cell malignancies.

Coordination of chromatin accessibility and mRNA transcription: which is first?

Finally, our data granularity, which captures the stepwise progression of closely related cells from lymphopoiesis (**Figure 1B** & **Figure 2A, B**), allowed us to interrogate the activation of the cCREs relative to the transcription of the gene(s) they regulate. As previously mentioned, an excellent correlation was observed between promoter accessibility and gene expression for the PAX5 gene (**Figure 1C, D**). However, our data showed that this strong association was not generalized; only part of the genes showed a significant and robust association with promoter OCR (**Figure S7A**). Nevertheless, when we explored the transition which reaches the highest transcriptional or chromatin accessibility change, a general correlation in the activation of the promoter and the gene transcription was observed (**Figure 5A, left**). We also observed a global delay in closing chromatin when gene expression decreases (**Figure 5A, right**). Yet, inspecting the cCREs, we found a different behavior. Here, a global delay in transcription was detected regarding chromatin accessibility at activation and deactivation (**Figure 5B** & **Figure S7B**). This result, in line with previous data on T cells[39], suggested that the enhancer elements precede transcription.

To further investigate the regulation of transcription activation, we explored the continuous dynamic of RNA and ATAC using single-cell technology in a human sample of CD34+ B cell precursors. Over 3,000 cells per omics were collected, including HSC, CLP, and pro-B cells (**Figure 5C** & **Figure S8**). The trajectory analysis drew the progression of the differentiation giving a continuous scenario to explore the RNA and ATAC coordination. Again, high coordination was observed within PAX5 transcription and promoter activity; both increased in pro-B cells (**Figure 5D** & **Figure 1C, D**). The delayed chromatin close was verified by exploring the promoter behaviour of the CFH gene, highly specific to HSCs (**Figure 5E, F**). Then, to confirm the early modifications of enhancer cCREs, we explored the behaviour of the CHD7 gene. CHD7 was upregulated in pro-B cells (and following stage[40]) but showed significant enhancer accessibility earlier (**Figure 5G, H**). To generalize this observation, we explored marker genes of HSC and CLP or pro-B cells (**Figure 5I**). A clear enrichment of signatures was observed matching with the

corresponding cell populations. Then, the enrichment of the cCREs associated with each gene set was explored. Interestingly, the promoter accessibility of genes enriched in HSC and CLP remained high in pro-B cells, according to the delay in closing. In contrast, the accessibility of promoters of pro-B cells was coordinated with the gene transcription (**Figure 5J**). Exploring the putative enhancer elements, earlier modifications of chromatin accessibility were observed for both signatures. For HSC and CLP signature, a decrease in chromatin accessibility was observed in CLP, preceding the reduction in gene expression. For the pro-B signature, chromatin accessibility increased at CLPs, preceding the rise of gene expression **(Figure 5K**). These multiomics single cell findings complemented our data obtained by bulk profiling, and fine-tuned the temporal relationship between chromatin accessibility and mRNA transcription during early B cell differentiation.

DISCUSSION

We addressed the challenge of characterizing the transcriptomic, chromatin accessibility, and the transcriptomic-chromatin interaction shifts during the B cell differentiation in humans. In addition, we aimed to elucidate the gene regulatory networks of B cell differentiation to complete what has been previously achieved in mice or mature stages of human B cell[9,11,12,15,41,42]. To this end, we first gathered - to our knowledge – the most extensive sample collection that includes the profiles for 8 cell subpopulations (HSCs, CLPs, pro-B cells, pre-B cells, Immature B cells, Transitional B cells, and Naive CD5+ and CD5- cells), matched omics of RNA-seq and ATAC-seq, for 13 healthy donors. The collection provides an unparalleled resource to robustly understand the system.

The deep profiling of transcriptomic and chromatin accessibility profiles delineates B cell differentiation (**Figure 1B**). The number of patients provided the statistical power for the ¡ characterizing key dynamics and identifying relevant genes and OCRs (**Figure S3D-F**). Our data confirms well-known biomarkers of early stages of B cell development[22,42], and provides a resource for a deeper study of each cell state (B-rex; https://brex.shinyapps.io/brex/). In addition, chromatin accessibility reveals a considerable heterogeneity within intronic and distal intergenic regions, which is in line to previous observations in mouse immune system and human hematopoiesis[9,21]. This increased variability may relate to their putative regulatory role as enhancer elements and points to a key function in B cell development (**Figure S3C**).

Since transcription and chromatin accessibility are not independent[43], we aimed to identify the CREs controlling B cell differentiation. Firstly, OCRs were differentiated according to genome location. OCRs located in promoter regions were assumed to be promoter elements, and OCRs located in intronic or distal intergenic regions were assumed to be possible enhancer elements[17] (**Figure S2A**). Then, gene-OCR pairing was conducted to the associated gene if the OCR was located in a promoter; however, for those defined as distal OCRs, we investigated all possible OCR-gene pairs within 1Mb of each gene TSS (**Figure S2B**). Such boundaries were derived from Yoshida, H et al.[9]. Although most of the significant gene-OCR pairs are derived from OCRs located 300Kb of TSS (**Figure S2F**), there's evidence of interactions that span over 1Mb[11,44], supporting the boundaries used.

This approach conducted on our large data set allowed us the robust identification of 62,559 OCR-gene pairs (**Figure S2B-E**). Although we are not defining causal relations, we provide a list of significant OCR-gene associations based on signal coordination that can target functional CREs[39]. The ideal scenario to validate identified cCREs would be to compare them with histone marks and chromatin conformation data from the same cell subpopulations and individuals. Instead, however, we made use of available public data. We made use of such data in two ways: first, to confirm that our cCREs are enriched by active histone marks and in the expected compartments of chromatin conformation (**Figure S2G, H**), and secondly, as a resource, to mark those OCR-gene pairs that are observed in chromatin conformation data from the closest system

available (Naïve B cells)[11,12,17]. For instance, we were able to validate 2,161 interactions that are characteristic of Naïve B cells in our data set and associated with genes highly expressed in late stages of B-cell differentiation as MS4A1 (CD20), CD79A/B, PAX5, and IKZF1 (**Figure S2I, J**). However, further efforts are needed to validate the identified cCREs in the early B-cell differentiation[45].

Defined the cCREs, we aimed to complete the regulatory information by annotating the OCRs with the candidate TFs binding to them. To this end, deep sequencing of ATAC-Seq data (over 100 million reads per sample) was performed[46]. This gives a promising scenario to identify TF footprints and confidently infer the TF binding motifs on chromatin-accessible regions. However, TF motif analysis is not trivial; the similarity between recognition sequences of TFs families' members makes TF motif finding challenging. Therefore, we used the non-redundant TF motif annotation from Vierstra[13] and the computation of the TFBS accessibility score that summarizes the enrichment of a TF over the entire set of OCRs for each sample[47]. Despite annotation challenges, when analyzing TFBS accessibility score through early B-cell differentiation (**Figure 2C**), activation of already known TFs for the differentiation and commitment of B cell lineage at corresponding differentiation stages are observed, such as EBF1, TCF3, GFI1, FOXO1, and PAX5 (highly correlated with the gene expression profiles (**Figure S3G**))[3,48]. Interestingly, based on the TFBS activity score, FOXD3 appears within the top variable TFs (**Figure S4C**). FOXD3 is a member of the fork head-box (FOX) family TFs related to neural crest development. Although scarce data on B cells and FOXD3 exists, it has been described as a pre-B cell-specific enhancer already primed in embryonic stem cells[49] and related to chronic lymphocytic leukemia (CLL) progression in mice[50]. Here, although we didn't detect high expression levels of FOXD3, and expression is not detected in public single-cell data[40], a relevant specificity of FOXD3 TF activity in pre-B cells in coordination with gene expression and promoter accessibility was observed (**Figure S4D, E**), pointing to a possible role of FOXD3 on B cell differentiation. Moreover, exploring the set of genes that FOXD3 can modulate within pre-B cells (n=397 genes), we identified two pathways related to the regulation of mitotic cell cycle phase transition enriched (GO:1901990 and GO:0044772) and enrichment of lymphoid leukemia and lymphoma disease ontologies, including B-cell lymphoma (DOID:1037, DOID:0060058, and DOID:707) (**Figure S4F**).

Overall, we described the activity (through TFBS activity score) of 169 TFs implicated in the B cell differentiation (B-rex), including well-known TFs at specific times, validating our data set and revealing new regulator candidates that require further investigation to decipher their roles.

By combining the OCR-genes pairs and TF-OCRs maps, we could reconstruct the GRN of early B-cell differentiation. Although a fraction of false positive edges is expected, we consider such network a powerful and very robust resource for hypothesis generation because (i) we used significant OCR-gene pairs derived from a large number of paired samples, (ii) we used footprinting allowed by deep sequencing, (iii) the GRN recovers the already known regulators, and (iv) a specific GRN for each cell state can be derived permitting us to investigate the TFs that switch-on and –off during differentiation (**Figure 2E, F**). Importantly, we observed that just using the GRNs, we can – similarly to what we observed in transcription and chromatin accessibility – reconstruct the differentiation trajectory. This is partly expected, as genes and OCRs were selected based on their differentiation, but for instance, the TF-TF GRNs, which have fewer features, still capture the system's dynamics (**Figure S5A**).

The recovered GRN shed light on understanding the B cell differentiation process and the transition to malignant diseases. Interestingly, we identified ELK3 as a new regulatory element critical in the early stages of differentiation (**Figure 2F, G**). This has been recently described as an oncogene[27–29] and related to ALL[24,26]. In addition, functional implications of ELK3 in pro-B cell expansion have been described[30], and its inhibition as a potential candidate therapy for tumor growth and metastasis in mice and in prostate cancer cells revealed[25,51,52]. Moreover, our data

highlight ETV6 as a candidate regulator of ELK3 (**Figure 3B, D** & **Figure S5D**), which could explain the increased expression observed in ETV6-RUNX1+ ALL cases[24,26] (**Figure 4E**). Also, our data reveal MLXIP, a TF associated with the proliferation control in B-ALL[32], highly active in pro-B and pre-B cells (**Figure 3E**) and a candidate regulator of EBF1 (**Figure 3A, C** & **Figure S5C**).

Aligning to a malignant scenario, our data allowed infer the putative cell origin of several ALL subtypes. Susceptibility points for malignancy transformation were highlighted in pre-B cells for MEF2D fusions, TCF3–PBX1, pro-B cells for ETV6–RUNX1, and HSCs for ZNF384 (**Figure 4C, D**). These observations align with the ETV6–RUNX1 and TCF3–PBX1 characteristics that mainly block B-cell differentiation and lead to the expansion of pre–B-cell clones[3]. Also, the association of ZNF384 fusions with HSC could explain the optimal treatment for this subtype of ALL patients, which currently is the allogenic-HSCT[53], and the enrichment of hematopoietic stem cell features in TCF3-ZNF384-positive patients[54]. Moreover, our data shed light on the functional implications of the susceptibility loci associated with ALL. Importantly, we identified rs17481869, described in ETV6-RUNX1[36], located in an OCR specific to pro-B cells (**Figure 4F**). We also found the rs4762284[26] in a closed chromatin region but revealed two proxies (rs2268508 and rs17736737) located within OCRs, one highly accessible in pro-B cells (**Figure 4G**). Other susceptibility loci were found within B-cell OCRs, revealing chromatin regions and loci that can trigger the ALL (**Table S4** & **Figure 4H-L**).

Finally, taking advantage of the data generated, we used a single-cell approach to investigate the coordination between the activation of the cCREs relative to the transcription of the gene(s) they regulate. Although single-cell data only cover three cell subpopulations studied, the analysis confirms our previous observations. Notably, a delay in promoter accessibility close was revealed when gene expression decreased. In contrast, earlier modifications of chromatin accessibility were observed for putative enhancer elements, pointing to enhancers as drivers of the transcription process (**Figure 5**).

In conclusion, the generated resource provides a complete atlas of regulatory elements governing B cell differentiation. The community can leverage this atlas as a hypothesis generator tool and to guide focused experiments to mechanistically understand the regulation of particular interactions in health and disease. To facilitate these parses, a user-friendly web interface for browsing has been designed (B-rex; https://brex.shinyapps.io/brex/).

MATERIAL & METHODS

All data and codes used in the project can be found on GitHub and Open Science Framework (OSF).

Human samples collection

Human bone marrow and peripheral blood samples from the same individuals were collected for 13 healthy donors (7 males and 6 females; median age of 21 [20-23]) (**Table S1**). Samples and data from donors included in the study were provided by the Biobank of the University of Navarra and were processed following standard operating procedures. The Clinical Research Ethics Committee of the Clinica Universidad de Navarra approved the study.

FACS sorting

MNC cells were purified from whole blood and bone marrow samples using Ficoll-Paque PLUS (GE Healthcare. Illinois, USA). After that, eight immune cell populations, representing lymphoid hematopoietic differentiation cascade along the B cell lineage from adult human bone marrow precursors and peripheral blood, were obtained from each sample. These cells were purified

following the sorting strategy shown in **Table S1** and **Figure S1** in a BD FACSAria II (BD Biosciences) and then analyzed by Bulk RNA-seq and Bulk ATAC-seq.

RNA-seq Data Generation

Bulk RNA-seq was performed following the MARS-Seq protocol adapted for bulk RNA-seq[55,56] with minor modifications. Briefly, cells were sorted in 100 µl of Lysis/Binding Buffer (Ambion), vortexed, and stored at -80ºC until further processing. Poly-A RNA was selected with Dynabeads Oligo (dT) (Ambion) and reverse-transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent) using poly-dT oligos carrying a 7 nt-index. Up to 8 samples with similar overall RNA content were pooled and subjected to linear amplification by IVT using HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs). Next, RNA was fragmented into 250-350 nt fragments with RNA Fragmentation Reagents (Ambion) and dephosphorylated with FastAP (Thermo) following the manufacturer´s instructions. Next, partial Illumina adaptor sequences were ligated with T4 RNA Ligase 1 (New England Biolabs), followed by a second reverse transcription. Full Illumina adaptor sequences were added during final library amplification with KAPA HiFi DNA Polymerase (Kapa Biosystems). RNA-seq libraries quality controls consisted of quantification with Qubit 3.0 Fluorometer (Life Technologies) and size profile examination using Agilent's 4200 TapeStation System. Libraries were sequenced in an Illumina NextSeq 500 in single-end at a sequence depth of 10 million reads per sample.

RNA-seq Quantification and Normalization

Reads quality was assessed with FastQC software (Babraham Institute). Low quality, short reads, and TruSeq3-SE adapters were trimmed using Trimmomatic[57] under the following parameters: ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30. Reads were then mapped to the December 2013 Homo sapiens high coverage assembly GRCh38 from the Genome Reference Consortium GRCh38 (hg38) using STAR [version 2.6.0a][58] and counted for each gene using htseq-count [59] with –m intersection-nonempty option and a GTF file from GRCh38 Ensembl annotation release 96, getting the information for a total of 58,884 genes. Genes with a minimum read count of 1 in all population replicates and an associated ensemble ID were retained (23,454 genes). A sample identified as an outlier has been removed (sample ID: D213-4). Counts were normalized under TMM normalization and voom transformation from edgeR r-package [v3.24.3] [60]. Correlation between samples was computed using Pearson correlation (**Figure S3A**). The summary of RNA-seq reads statistics can be found in **Table S2.**

RNA-seq Data analysis

Principal Component Analysis (PCA) was used for data exploration (**Figure 1B**). Differential expression analysis (DEA) was performed using limma [v3.38.3] and edgeR [v3.24.3] under two different approaches: 1) Identify biomarker genes for each cell subpopulation (all possible pair combinations); 2) Identify differential expression genes (DEGs) between transition states (HSC vs. CLP, CLP vs. pro-B, pro-B vs. pre-B, etc.). Genes with an adjusted p-value ≤ 0.05 (Benjamini-Hochberg) and absolute fold-change ≥ 1.5 were considered significant. A clustering analysis was performed to identify the different expression patterns within the differentially expressed genes over transitions (second approach) using a model-base clustering based on parameterized finite Gaussian mixture models (mclust r-package [v5.4.5] under the following function and parameter: Mclust (modelName="EII")[61]) (**Figure S3D**). Gene set enrichment analysis of GO terms was performed using ClusterProfiler r-package [v4.6.0][62]. The fold-change of DEGs at different transitions was given as input to the *compareCluster* function, and the top 10 enriched terms were plotted using an in-house *treeplot* function (**Figure S3E**).

ATAC-seq Data Generation

Accessible chromatin mapping was performed using FAST-ATAC-seq [21] with minor modifications. Briefly, 10,000 cells were sorted in 1X PBS, 0.05% BSA and pelleted by

centrifugation at 500 rcf for 5 min at 4 °C with low acceleration and brake settings in a pre-cooled swinging-bucket rotor centrifuge. All the supernatant but 5 µL was removed. Next, 25 µL of transposase mix (15 µL of 2x TD buffer (Illumina); 1 µL of TDE1 (Illumina), 0.25 µL of 5% digitonin (Promega); 8.75 µL of nuclease-free water) were added to the cells, and the pellet was disrupted by pipetting. Transposition reactions were incubated at 37 °C for 30 min in an Eppendorf ThermoMixer with shaking at 450 rpm. Reactions were stopped at 4 °C for 5 min. To release tagmented DNA, samples were incubated at 40 °C for 30 min with 5 µL of clean-up buffer (900 mM NaCl (Sigma), 30 mM EDTA (Millipore), 2 µL of 5% SDS (Millipore) and 2 µL of Proteinase K (NEB)). DNA was purified using a 2X SPRI beads cleanup kit (Agencourt AMPure XP, Beckman Coulter). Two sequential PCRs were performed using KAPA HiFi DNA Polymerase and customized Nextera PCR indexing primers (IDT) [63]. The conditions of the first PCR were: 5 min at 72 °C and 2 min at 98 °C followed by 9 cycles of 20 secs at 98 °C, 30 secs at 63 °C, and 1 min at 72 °C. Depending on the library concentration in this first PCR, a second PCR (2 min at 98 °C followed by 4 to 6 cycles of 20 secs at 98 °C, 30 secs at 63 °C, and 1 min at 72 °C) was performed aiming for a library concentration in the range of 2 to 10 ng/µL. PCR products were purified using a 2X SPRI beads cleanup. Libraries were quantified, and their size profiles were examined as described above. Sequencing was carried out in two batches, the first one in an Illumina NextSeq 500 using paired-end, dual-index sequencing (Rd1: 38 cycles; Rd2: 38 cycles; i7: 8 cycles; i5: 8 cycles) at a depth of 10 million reads per sample and second one in a HiSeq2500, v4 reagents using paired-end, dual-index sequencing (Rd1: 125 cycles; Rd2: 125 cycles; i7: 8 cycles; i5: 8 cycles) at a depth of 100 million reads per sample.

ATAC-seq Quantification and Normalization

Reads quality was assessed with FastQC software (Babraham Institute). After trimming NexteraPE-PE adaptor sequences and low-quality reads and cutting the reads from the second batch to 38 bps using trimmomatic[57] under the following parameters: CROP:38 ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:15. Reads were then mapped to hg38 reference genome (GRCh38 Ensembl annotation release 96) using BWA-MEM [64]. Unmapped, unknown, random, ChrM mapping, duplicate reads, and reads matching to black list (the blacklist includes problematic regions of the genome[65]) were filtered out using samtools view -b -F 4 -F 1024 (samtools version 0.1.19), and Picard Tools (Picard MarkDuplicates, http://broadinstitute.github.io/picard). BAM files from re-sequenced samples were merged using samtools merge. The summary of ATAC-seq reads statistics can be found in **Table S2**. Paired-end reads were used to determine those regions with significant reads coverage in contrast to the random background (regions known as peaks or Open Chromatin Regions; OCRs) using MACS2 (function *callpeak*) with the following parameters: -g hs -f BAMPE -q 0.01 --nomodel --shift -100 --extsize 200 -B –SPMR (https://github.com/taoliu/MACS) [66]. A minimum OCR size of 20 bps was set as a threshold. To get the consensus OCRs for each cell subpopulation, we found the genome regions which are covered by at least n-1 of the sets of peaks (n being the total number of samples for a specific cell subpopulation), and then we merged the nearby peaks with a min gap width of 31. To get the consensus OCRs in the whole population, we get the regions in the genome where coverage is at least in 1 of the consensus OCRs from the different cell subpopulations and then merged the nearby peaks with a min gap width of 31, resulting in 105,530 ATAC-seq OCRs. These processes were done using the rtracklayer [v1.58.0] and GenomicRanges [v1.50.2] r-packages[67,68].

BigWig files were generated for each BAM file using the *bamCoverage* function (--binSize 20 --normalizeUsing RPKM --effectiveGenomeSize 2913022398) from deepTools [v3.3.1] [69]. Then, a merged BigWig for each cell subpopulation was obtained through *bigWigMerge* from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.x86_64/). Also, BigBed files were generated from consensus BED files of each cell subpopulation using *bedToBigBed* from UCSC. These data were deposited in a public custom track from UCSC (https://genome.ucsc.edu/s/TransBioLab/bcell_hg38).

Using the whole consensus OCRs, we compute the count matrix using the *multicov* function from bedtools [version 2.25.0] [70]. Those OCRs without a minimum of 10 counts for at least one sample and 15 total counts were excluded (155 OCRs), retaining 105,359. Then, we normalize the data using the TMM method from edgeR r-package [v3.30.3] [60]. One of the samples analyzed was identified as an outlier and excluded (sample ID: D215-5). Correlation between samples was computed using Pearson correlation (**Figure S3A**).

ATAC-seq Data analysis

Principal Component Analysis (PCA) was used for data exploration (**Figure 1B**). OCRs were mapped to genomic locations (promoter, intronic, distal intergenic, etc.) using the ChIPseeker r-package [v1.18.0] [71] under the following arguments: Region Range of TSS as +/- 1000 bps, gene level, "TxDb.Hsapiens.UCSC.hg38.knownGene" TxDb object [v3.4.0] and "org.Hs.eg.db" annotation package [v3.7.0] (**Figure S2A**). The Gini index was computed for each OCR to measure "chromatin inequality" (**Figure S3B**). A Gini index of 0 represents perfect equality and similar chromatin accessibility in all cell subpopulations; while a Gini index of 1 represents perfect inequality, all the accessibility OCRs come from a specific cell subpopulation, and the others show no accessibility. To compute the Gini index, the maximum chromatin accessibility observed from the normalized count matrix was assigned for each cell subpopulation. Then, the *gini* function from REAT r-package [v3.0.2] was applied [72]. Differential accessibility analysis (DAA) was performed to identify differential accessibility regions between transition states (HSC vs. CLP, CLP vs. pro-B, pro-B vs. pre-B, etc.) using a quasi-likelihood negative binomial generalized log-linear model (glmQLFit) from edgeR [v3.24.3] [60,73]. OCRs with an adjusted p-value ≤ 0.01 (Benjamini-Hochberg) and absolute fold-change ≥ 2 were defined as significant. A clustering analysis was performed to identify the different accessibility patterns within the differentially accessible regions under a model-base clustering based on parameterized finite Gaussian mixture models (mclust r-package [v5.4.5] under the following function and parameter: Mclust(modelName="EII"))[61] (**Figure S3F**).

OCRs association to target genes (cis-regulatory elements)

To determine the cis-regulatory elements (CREs) of a gene, we follow two approaches: 1) For promoter OCRs, we directly annotate OCRs to genes based on TSS distance (<1kb), and 2) For intronic or distal intergenic OCRs, we explored the association between chromatin accessibility and gene expression. First, we identified all intronic or distal intergenic OCRs for each expressed gene within a window of ±1Mb from the gene's TSS. Second, the Spearman correlation was computed across all samples (59 samples; 8 cell subpopulations) for each gene-OCR pair (**Figure S2B**). Bonferroni correction was used to adjust for multiple testing. Gene-OCR pairs with adjusted p-value < 0.05 were considered significant. A complete list of pairs (62,559) is available in OSF.

OCRs association with transcription factors (trans-regulatory elements)

To annotate OCRs to putative transcription factor binding motifs, we first identify the footprints within our OCRs. For each sample, we recognize the footprints using two algorithms, HINT[74] and Wellington [75], using as input the corresponding sample BAM file and BED file from the corresponding cell subpopulation. For the footprints identified by HINT, we filtered out those footprints with less than 20 counts. Then, we get the union of the footprints identified for each sample through both strategies. Following, we compute the consensus footprint for each cell subpopulation following a similar strategy as for OCR consensus. To get the consensus footprint for each cell subpopulation, we found the genome regions covered by at least n-1 of the sets of regions (n being the total number of samples for a specific cell subpopulation). Finally, to get the consensus footprint in the whole population, we get the regions in the genome where coverage is at least in 1 of the consensus footprint from the different cell subpopulations; then, those with a width less than 6bps were removed, resulting in a total of 380,922 ATAC-seq footprints. These processes were done using the rtracklayer [v1.58.0] and GenomicRanges [v1.50.2] r-packages[67,68]. The identified footprints were located in 72,771 OCRs (69%).

Identified the footprints, we want to annotate the putative transcription factor binding motif to each footprint. To do that, we used the Vierstra database[13]. First, we generate a reduced version of the Vierstra database, getting those annotations falling within our OCRs regions using TABIX[76]. Then, we annotated at the archetype level (v1.0) to reduce motif redundancy and overlap, resulting in 30,258,213 region-motif annotations and 286 motif archetypes. The overlap between footprints and motif archetypes was explored, defining a minimum overlap of 90% of archetype width. Finally, footprints were linked to the original OCR to get the OCR-Archetype matrix, resulting in a matrix 105,530 x 282.
To translate archetypes to TFs, we define the following strategy. First, we compute the aggregated motif score (TFBS accessibility score) for each archetype using chromVAR r-package [v1.5.0][47] with the default parameters. Then, to identify the TFs most likely associated with an archetype, we compute, for each archetype, the Spearman correlation between the TFBS score of the archetype and the RNA expression of each TFs clustered in that archetype. Correction for multiple comparisons was done using the Benjamini-Hochberg method. Moreover, the TFBS score variability and the coefficient of variation of TF gene expression were computed. We determine the TF-archetype association when: 1) Significant variation for both TFBS score and TF gene expression and significant and positive correlation is observed (adjusted p-value <0.05 and rho>0), 2) No considerable variation is observed for TFBS score (<1Q TFBS variability) or TF gene expression (<1Q RNA coefficient of variation), independently of the correlation result, assuming that correlation will not be meaningful due to the low variability of the variables assessed. Finally, we obtained a list of archetype-TF associations that allowed us to translate archetypes to TFs and get the OCR-TF matrix (105,530 x 275) with 60,881 OCRs linked at least to one TF motif.

Generating the Gene Regulatory Networks of B cell differentiation

Based on the different associations of OCRs with genes (cis-regulation; 62,559 links) and TFs (trans-regulation; 2,299,328 links), we were able to define the gene regulatory network (GRN) for early B cell differentiation. To do that, from the defined OCR-gene and OCR-TF pairs, we first keep those interactions derived from the DEGs (differential expression at least in one transition), their promoter regions, and the associated enhancer elements with differential accessibility (differential chromatin accessibility at least in one transition); under the assumption that those genes and OCRs changing across cell subpopulations are going to be the ones governing the B cell differentiation process. Then, having the putative TF-OCR-gene interactions (n=1,151,024), the correlation between TF-OCR and TF-gene was computed using Spearman correlation. To reduce spurious and meaningless TF-OCR-gene interactions, we excluded from the GRN those where: 1) TF showed no differential expression (n= 469,073), 2) TF-gene association was not significantly correlated (Spearman p-value >0.05; n=186,837), 3) Meaningless interactions including (a) positive correlation for TF-gene and TF-OCR but negative correlation for TF and gene, (b) negative correlation for TF-gene and TF-OCR but positive correlation for TF and gene, (c) positive correlation for TF-gene and TF-OCR but negative correlation for OCR-gene, (d) negative correlation for TF-gene, TF-OCR and OCR-gene, and (e) positive correlation for TF-gene and OCR-gene but negative for TF-OCR (n=117,846). Finally, 377,268 TF-OCR-gene interactions were retained to reconstruct the GRN of the B cell differentiation process, involving 169 TFs, 7,310 genes, and 16,074 OCRs.
Then, the specific GRN for cell subpopulation was computed based on the chromatin accessibility and gene expression for each cell subpopulation. For example, for HSCs, the GRN captures those TF-OCR-gene interactions where genes are highly expressed in that cell subpopulation according to gene expression patterns (**Figure 2A**, genes included in brown clusters) and a median of normalized chromatin accessibility ≥ 1 (defining accessible OCR) is observed. Network visualization was done with Cytoscape [v3.7.2][77].

Computing aggregated TF motif scores and association with TF gene expression

To explore the accessibility of TFs from our GRN, we compute the aggregated motif score (TFBS accessibility score) for each TF using chromVAR r-package [v1.5.0] [47] with the default parameters (**Figure 2C**). Then, we compute the correlation between TFBS accessibility score (raw deviations) and RNA expression (normalized data) for each TF using Spearman correlation (**Figure S3G**). Multiple comparisons were corrected using the Benjamini-Hochberg method. Adjusted p-value < 0.05 was defined as significant.

Topological properties and network distances

According to the structure of our GRNs, different network properties were computed using the igraph r-package [v1.3.5][78]. Node betweenness (the amount of influence a node has over the flow of information in a graph through the shortest paths calculation between all pairs of nodes) and degree (the total number of edges adjacents to a node) were evaluated for the complete set of GRNs with weighted edges according to the Spearman's coefficient of correlation between TF and gene (**Figure 2E**). Moreover, Pearson correlation was computed to compare the networks between them, and multidimensional scaling was used for visualization (**Figure 2D**).

Public ChIP-Seq Histone marks

ChIP-Seq data from ENCODE project (https://www.encodeproject.org/)[79] for H3K4me1, H3K4me3, and H3K27ac on B cells, naive B cells, and common myeloid progenitors CD34+ was downloaded (**Table S1**). To get the consensus regions for each cell subpopulation and histone mark, we found the genome regions which are covered by at least n-1 of the sets of peaks (n being the total number of samples for a specific cell subpopulation), and then we merged the nearby peaks with a min gap width of 31. To get the consensus regions for a histone mark in the whole population, we get the regions in the genome where coverage is at least in 1 of the consensus histone marks from the different cell subpopulations and then merged the nearby peaks with a min gap width of 31, resulting in 166,295 H3K4me1 regions, 37,952 H3K4me3 regions, and 159,079 H3K27ac regions. These processes were done using the rtracklayer [v1.58.0] and GenomicRanges [v1.50.2] r-packages [67,68]. Then, the overlap between our OCRs and histone marks was assessed using the *findOverlaps* function (minoverlap=6) of GenomicRanges (**Figure S2G**). The chi-Squared test was used to test the histone enrichment throughout different OCR types. A p-value <0.05 was considered significant.

Public Hi-C data

Hi-C public data of naive B cells were obtained from EGAD00001006486[12].

The valid reads of Hi-C experiments were processed with TADbit version 1.1 [80]. Hi-C data were normalized using the *oned* function of dryhic r-package [v0.0.0.9100] [81] at 100 kb of resolution to remove known experimental biases, a method that controls for the presence of abnormal karyotypes in cancer samples. Topologically associating domains (TADs) were detected at 40 kbp resolution using the TopDom algorithm [82]. The significant Hi-C interactions were called with the Fit-Hi-C tool [v1.1.3] [83,84], binned at 10 kb of resolution and with the default thresholds; p-value< 0.001 and q-value<0.001.

To study the compartmentalization of the genome in naive B cells, we used merged biological replicates resulting in interaction Hi-C maps with around 300 million valid reads each. Briefly, normalized chromosome-wide interaction matrices at 100 kb resolution were transformed into Pearson correlation matrices. These correlation matrices were then used for PCA, for which the first eigenvector (EV) normally delineates genome segregation. The multi-modal distribution of the EV coefficients from the B cells dataset was modeled as a Gaussian mixture with three components (k = 3). To estimate the mixture distribution parameters, an Expectation-Maximization algorithm using the *normalmixEM* function from the mixtools r-package [v1.2.0] was applied [85]. A Bayesian Information Criterion (BIC) was computed for the specified mixture models of clusters (from 1 to 10) using the *mclustBIC* function from mclust r-package [v5.4.9] [61]. Three underlying structures were defined; alternative compartmentalization into A-type (with the most

positive EV values), B-type (with the most negative EV values), and I-type (an intermediate-valued region with a different distribution) compartments. Two intersection values were defined at the intersection points between two compartments and used as standard thresholds to categorize the data into the three different compartments; that is: the A-type compartment was defined for EV values >= +0.1388, the I-type compartment was defined for EV values between +0.1388 and −0.01955, and the B-type compartment was defined for EV values between <= −0.01955.
Then, we check the overlap between our OCRs and the compartments identified using the *findOverlaps* function (default parameters) of GenomicRanges r-package [v1.50.2] (**Figure S2H**). When an OCR overlaps with two compartments, we assign it to the compartment with a higher overlap. The chi-Squared test was used to test the compartment enrichment throughout different OCRs. A p-value <0.05 was considered significant.

Public promoter capture Hi-C interaction data

Detected interactions using promoter capture Hi-C (PCHiC) technology from Javierre et al.[11] were downloaded. From the PCHiC peak matrix, we select all those interactions with a CHiCAGO score ≥ 5 in the naive B cell population, resulting in 197,442 interactions. The coordinates of the interacting regions were translated from GRCh37 to GRCh38 using liftOver r-package [v1.18.0]. Then, we check the overlap between our interactions (OCR-gene associations) and the interactions they identified using the *findOverlaps* function (minoverlap=6) of GenomicRanges r-package [v1.50.2] (**Figure S2I, J**).

Public RNA-Seq data from BCP ALL patients

Transcriptomic data from BCP ALL patients (leukemic cells from bone marrow aspirates) were obtained from the European Genome-phenome Archive. The RNA-Seq data (BAM files) from EGAD00001002151, EGAD00001002692, and EGAD00001002704 datasets were downloaded after the corresponding access request. Only data from pediatric individuals belonging to a well-known molecular subgroup[34], ranging from G1 to G8 (MEF2D fusions, TCF3–PBX1, ETV6–RUNX1/-like, DUX4 fusions, ZNF384 fusions, BCR–ABL1/Ph-like, Hyperdiploidy, KMT2A fusions) were included in the analysis, resulting in 356 patients (**Table S1**).
As BAM files were mapped to GRCh37, we first obtained the Fastq files using the *bamtofastq* function from bedtools [version 2.25.0] [70]. Next, the read quality was assessed with FastQC software [version 0.11.7] (Babraham Institute). Low quality, short reads, and TruSeq3-SE adapters were trimmed using Trimmomatic [version 2.6.0a] [57] under the following parameters: ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75. Reads were then mapped to the December 2013 Homo sapiens high coverage assembly GRCh38 from the Genome Reference Consortium GRCh38 (hg38) using STAR [version 2.6.0a][58] and counted for each gene using htseq-count[59] with –m intersection-nonempty option and a GTF file from GRCh38 Ensembl annotation release 96, getting the information for a total of 58,884 genes. Genes with a minimum read count of 1 in all population replicates and an associated ensemble ID were retained (25,914 genes). Next, the batch correction was done using the *ComBat_seq* function from sva r-package [v3.38.0] [86,87]. Counts were normalized under TMM normalization and voom transformation from edgeR r-package [v3.32.1][60]. Multidimensional scaling (MDS) was used for data exploration. Differential expression analysis (DEA) of leukemic subgroups was performed using limma. Those genes with low coefficient of variation (<25% cv) were filtered out before DEA to win statistical power. Genes with an adjusted p-value ≤ 0.01 (Benjamini-Hochberg) and absolute fold-change ≥ 1 at least in 4 contrasts were defined as significant (subgroup biomarkers). The subgroup biomarkers list can be found in **Table S4**.

Over-representation analysis in public BCP ALL data

The over-representation analysis (ORA) of the gene sets from BCP ALL subgroups (**Table S4**) was performed in 1) the biomarkers of the different B-cell population (**Figure 4B**) and 2) the core genes derived from the mutated TF's downstream regulation (**Figure 4C**), using the *enricher* function of clusterProfiler r-package [v3.18.1][62]. The gene ratio and –log10 from the adjusted p-

value under the Benjamini-Hochberg method were computed. The ORA was also applied to explore the enrichment of 711 binding sites defined as the highest affinity binding sites for ETV6-RUNX1 fusion under ChiP-Seq analysis[35] in the consensus OCRs for each cell subpopulation (**Figure 4D**).

GWAS for ALL

The gwascat r-package [v2.30.0], which provides functionalities to work with the contents of EMBL-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/) [88], was used to explore the single nucleotide polymorphisms (SNPs) associated with ALL. From the entire database containing 474,151 records, 256 records related to disease term involving the string "acute lymphoblastic leukemia" were selected (**Table S4**). The overlap between selected records and the entire list of our OCRs (n=105,530) was assessed using the *findOverlaps* function of GenomicRanges r-package [v1.50.2].

Single-cell RNA-seq (scRNA-Seq) sample

An additional human bone marrow sample (female, 21 years old) was included in the study for single-cell analysis. The sample was sorted for hematopoietic lineage (sorting strategy: CD34+ CD38- CD3- CD64- CD56- CD90+), and viable cells were obtained and divided to perform both scRNA-Seq and scATAC-seq analysis.

Single-cell RNA-seq (scRNA-Seq) Data Generation

The transcriptome of the hematopoietic cells from the bone marrow was examined using NEXTGEM Single Cell 3' Reagent Kits v2 (10X Genomics), according to the manufacturer's instructions. Between 17,600 and 20,000 cells were loaded at a concentration of 700-1,000 cells/µL on a Chromium Controller instrument (10X Genomics) to generate single-cell gel bead-in-emulsions (GEMs). In this step, each cell was encapsulated with primers containing a fixed Illumina Read 1 sequence, a cell-identifying 16 nt 10X barcode, a 10 nt Unique Molecular Identifier (UMI), and a poly-dT sequence. Upon cell lysis, reverse transcription yielded full-length, barcoded cDNA. This cDNA was then released from the GEMs, PCR-amplified, and purified with magnetic beads (SPRIselect, Beckman Coulter). Enzymatic Fragmentation and Size Selection were used to optimize cDNA size before library construction. Fragmented cDNA was then end-repaired, A-tailed, and ligated to Illumina adaptors. Final PCR amplification with barcoded primers allowed sample indexing. Library quality control and quantification were performed using Qubit 3.0 Fluorometer (Life Technologies) and Agilent's 4200 TapeStation System (Agilent), respectively. Sequencing was performed in a NextSeq500 (Illumina) (Read1: 26 cycles; Read2:55 cycles; i7 index: 8 cycles) at an average depth of 60,226 reads/cell.

Single-cell RNA-seq Quantification, Normalization, Clustering, and Annotation

The binary base call (bcl) sequence file format obtained on the 10x Genomics Chromium sequencing platform was transformed into fastq format using mkfastq from Cell Ranger 3.0.1 [55]. The count matrix used for downstream analysis was generated using a count pipeline from Cell Ranger 3.0.1 with GRCh38 reference transcriptome provided by 10X Genomics.

Following Seurat guidelines[89–92], a four-step cell and feature quality filtering were applied. First, based on lower (0.1) - upper (0.9) bound quantile for "number of features" and "counts per cell", followed by a minimum feature per cell filtering allowing cells present in at least 201 genes and a feature removing strategy based on "minimum features per cell", retaining only genes present in at least 4 cells. Finalizing a cell cut-off base on maximum mitochondrial percentage was applied, allowing cells containing less than 5% mitochondrial gene content. Feature selection criteria have been used for downstream analysis based on most variable genes, using 3,000 top genes. Normalization of the gene expression measurements for each cell was done by multiplying the total expression by a scale factor (10,000) and log-transforming it. Feature subspaces were obtained from most variable genes using principal component analysis (PCA) using 30

components. Clustering was computed using the Louvain algorithm over principal components subspace with a resolution value of 0.6. For data visualization, UMAP was used, computing it over all PCA components obtained.

To annotate cell identity, our bulk RNA-seq data was used as a reference for training the SingleR classifier to later scRNA-seq cells classification [93]. The following considerations have been taken for SingleR input preparation. First, marker genes with adjusted p-value < 0.01 and absolute fold-change greater than 1.5 were selected for each pair of annotated cell subpopulations to get the most representative genes involved in transitions of RNA-seq cell subpopulations. Then, scRNA-seq cluster markers were computed considering only genes in at least 25% of cells in each group and filtering based on adjusted p-value < 0.01 and logFC of 0.25. Overlapping genes obtained from both analyses were then filtered based on their belonging to the top 1,500 genes represented in more cells in scRNA-seq. This enhanced set of genes well detailed in most scRNA-seq cells that captures the transitions from RNA-seq and relevance for cluster identification in scRNA-seq was finally used from RNA-seq expression matrix as input. Already contained annotations in RNA-seq (HSC, CLP, pro-B, pre-B, Immature B, etc.) were used as sample labels for training SingleR. Afterward, a trained SingleR model was used to classify single cells into each bulk cell subpopulation label. Finally, scRNA-seq was filtered based on only HSC, CLP, and pro-B cell subpopulations removing the rest of the cells not belonging to the B cells differentiation process in this study. For signatures computation, the normalized and log2 transformed expression values of the signature subset features are summed up for each cell and divided by the total expression values from all feature space sum of the corresponding cell. Finally, values for each cell are multiplied by a factor of 100 (**Figure S8**).

Single-cell ATAC-seq Data Generation

The scATAC experiments were performed with the Chromium Next GEM Single Cell ATAC kit v1.1 from 10X Genomics following the manufacturer's instructions. For samples with 100,000 cells or less, the nuclei isolation is done following the 10X Genomics demonstrated protocol for low cell input. Briefly, cells are collected in PBS+0.04% BSA at 4ºC, and after that, cells are counted in a cell counter, and centrifugation is done to concentrate them in 50ul of PBS+0.04% BSA. After that, 45ul of supernatant is removed, and 45ul of chilled and freshly prepared lysis buffer is added. After 5 minutes of incubation on ice, 50ul of wash buffer is added. After that, centrifugation is done to remove 95ul of the supernatant without disrupting the pellet. Additional centrifugation is done with 45ul of diluted nuclei buffer to remove the supernatant. Isolated nuclei are resuspended in 7ul of chilled diluted nuclei buffer. Two of these microlitres are used to determine the cell concentration using a cell counter. The other 5ul are used immediately to generate a bulk transposition reaction following the manufacturer's instructions and conditions. In the transposase reaction, the transposase enters the nuclei and preferentially fragments the DNA in open regions of the chromatin. In this step, adapter sequences are added to the ends of the DNA fragments. After transposition, GEMs are generated in the Chromium controller (10x Genomics). Upon GEM generation, the gel bead is dissolved. Oligonucleotides containing an Illumina P5 sequence, a 16nt 10X barcode, and a read1 sequence are released and mixed with DNA fragments and master mix. After incubation, GEMs are broken, and pooled fractions are recovered. P7, a sample index, and read2 sequence are added during library construction via PCR. Library quality control and quantification were performed using Qubit 3.0 Fluorometer (Life Technologies) and Agilent's 4200 TapeStation System. Bioanalyzer 2100 (Agilent). Sequencing was performed using a NextSeq500 (Illumina) (Rd1 50 cycles, i7 Index 8 cycles, i5 index 16 cycles, and Rd2 50 cycles) to an average depth of 14,900 fragments/cell.

Single-cell ATAC-seq Quantification, Normalization, Clustering, and Annotation

The bcl file format obtained on the 10x Genomics Chromium sequencing platform was, in this case, transformed into fastq format using mkfastq from Cell Ranger ATAC 1.1.0 [10]. The count matrix used for downstream analysis was generated similarly using count pipeline from Cell

Ranger ATAC 1.1.0. The Cell Ranger ATAC GRCh38 reference file provided by 10X Genomics was used. Following a schema similar to the one used in the scRNA-seq analysis, scATAC-seq data was also preprocessed. In this case, a combined Seurat and Signac guideline was used [89–92,94]. In addition to the same cell filtering applied before in scRNA-seq, additional cell filtering was applied over the scATAC-seq dataset. The ratio of mononucleosomal to nucleosome-free fragments per cell was calculated by filtering cells with a ratio greater than 4. The transcription start site (TSS) enrichment score for each cell as defined by ENCODE was also computed considering cells with greater values than 2. The total number of fragments in peaks for each cell was used to retain cells that have greater than 3,000 and smaller than 100,000 number of fragments in peaks. The percent of reads in peaks was also used to filter cells, considering only those that have more than 40 percent of reads in peaks for each cell. Blacklist regions reference was used to remove those cells with a ratio of reads falling in these regions greater than 0.025[65].

Data were normalized using the frequency-inverse document frequency (TF-IDF) method[91]. Reducing the features to a list of the most variable peaks is an option that may reduce the performance. Instead, we used the entire feature space in this case. Reduced feature subspaces were computed over all peak features space using singular value decomposition (SVD), providing latent semantic indexing (LSI) as latent space with 50 components. Clustering outcomes were obtained by applying the Louvain algorithm over the LSI components subspace, not considering the first component because of its deep correlation to sequencing depth. UMAP was applied for 2-D visualization employing the same LSI components used for clustering with a resolution value of 0.6.

Aiming to perform label transfer from scRNA-seq using Seurat protocols, sc-ATAC-seq preprocessed data was used for generating scRNA-simulated data (activity expression); to this end, the Signac activity estimation approach was used. An upstream of 2,000 base pairs was used for “peak to gene relation” estimation. Single cell ATAC-seq activity matrix was used to find anchors against the reference scRNA-seq expression matrix. Finally, cell annotation was transferred by classifying each cell from scATAC-seq using a computed anchor set over LSI components subspace (except the first component). Re-assignation was done at the clustering level, assigning cells from the same cluster to the most enriched cell subpopulation predicted to exclude inconsistencies in cell assignation.

To make bulk and single-cell ATAC-seq comparable, a re-quantified version of the scATAC-seq original raw matrix was conducted to generate a new feature x cell raw count matrix with same OCRs from bulk as feature space using the *FeatureMatrix* function from Signac package[94]. The new scATAC-seq matrix contains recalculated count values base on scATAC-seq genomic fragments according to bulk ATAC-seq feature space (**Figure S8**). For signatures computation, the normalized and log2 transformed expression values of the signature subset features (OCRs) are summed up for each cell and divided by the total expression values from all feature space sum of the corresponding cell. Finally, values for each cell are multiplied by a factor of 100.

Pseudotime analysis

For the scRNA and scATAC data, a pseudotime analysis was performed through Monocle3 [version 1.0.0][95] and Cicero [version 1.3.6][96] r-packages, respectively, with default parameters. The pseudotime values were scaled from 0 to 100.

DATA AND SOFTWARE AVAILABILITY

The EGA accession number for RNA-seq and ATAC-seq data reported in this paper is EGAS00001007296. Processed data, results, and codes used in the project can be found on GitHub, OSF, and in a user-friendly web interface for browsing (B-rex; https://brex.shinyapps.io/brex/).

ACKNOWLEDGEMENTS

This work was supported by the Instituto de Salud Carlos III and co-financed by ERDF A way of making Europe (PI17/00701, and PI20/01308), CIBERONC (CB16/12/00489) and RICORS TERAV (RD21/0017/0009); Gobierno de Navarra (AGATA 0011-1411-2020-000010/0011-1411-2020-000011; Fundación La Caixa (GR-NET NORMAL-HIT HR20-00871); and Cancer Research UK [C355/A26819], FC AECC and AIRC under the Accelerator Award Program. The Multiple Myeloma Research Foundation Networks of Excellence 2017 Immunotherapy Program Grant Award, the International Myeloma Foundation (Brian van Novis), the Paula and Rodger Riney Foundation.

N.P.P was funded by a Juan de la Cierva-formación fellow (FJC2019-042304-I) from the Spanish Ministry of Science and Innovation (MCIN) and by a Ramón y Cajal fellow (RYC2021-032197-I) from the MCIN/AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRTR.

This study makes use of data generated by the St. Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project (EGAD00001002692 and EGAD00001002704), and data generated by IDIBAPS (EGAD00001006486).

We particularly acknowledge the healthy donors for their participation in this study, and the Biobank of the University of Navarra for its collaboration.

REFERENCES

1. Somasundaram, R., Prasad, M. A. J., Ungerbäck, J. & Sigvardsson, M. Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood* **126**, 144–152 (2015).
2. Nutt, S. L. & Kee, B. L. The Transcriptional Regulation of B Cell Lineage Commitment. *Immunity* **26**, 715–725 (2007).
3. Fischer, U. *et al.* Cell Fate Decisions: The Role of Transcription Factors in Early B-cell Development and Leukemia. *Blood Cancer Discov.* **1**, 224–233 (2020).
4. Somasundaram, R. *et al.* EBF1 and PAX5 control pro-B cell expansion via opposing regulation of the Myc gene. *Blood* **137**, 3037 (2021).
5. Treiber, T. *et al.* Early B Cell Factor 1 Regulates B Cell Gene Networks by Activation, Repression, and Transcription- Independent Poising of Chromatin. *Immunity* **32**, 714–725 (2010).
6. Medvedovic, J., Ebert, A., Tagoh, H. & Busslinger, M. Pax5: A Master Regulator of B Cell Development and Leukemogenesis. *Adv. Immunol.* **111**, 179–206 (2011).
7. Yoshida, T. & Georgopoulos, K. Ikaros fingers on lymphocyte differentiation. *Int. J. Hematol.* **100**, 220–229 (2014).
8. Ikawa, T., Kawamoto, H., Wright, L. Y. T. & Murre, C. Long-Term Cultured E2A-Deficient Hematopoietic Progenitor Cells Are Pluripotent. *Immunity* **20**, 349–360 (2004).
9. Yoshida, H. *et al.* The cis-Regulatory Atlas of the Mouse Immune System. *Cell* **176**, 897-912.e20 (2019).
10. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol. 2019 378* **37**, 925–936 (2019).
11. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).
12. Vilarrasa-Blasi, R. *et al.* Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. *Nat. Commun.* **12**, 651–651 (2021).
13. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nat. 2020 5837818* **583**, 729–736 (2020).

14. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
15. Lee, R. D. *et al.* Single-cell analysis identifies dynamic gene expression networks that govern B cell development and transformation. *Nat. Commun.* **12**, (2021).
16. Decker, T. *et al.* Stepwise activation of enhancer and promoter regions of the B cell commitment gene Pax5 in early lymphopoiesis. *Immunity* **30**, 508–520 (2009).
17. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nat. 2020 5837818* **583**, 699–710 (2020).
18. Gates, L. A., Foulds, C. E. & O'Malley, B. W. Histone marks in the 'drivers seat': functional roles in steering the transcription cycle. *Trends Biochem. Sci.* **42**, 977 (2017).
19. van Mierlo, G., Pushkarev, O., Kribelbauer, J. F. & Deplancke, B. Chromatin modules and their implication in genomic organization and gene regulation. *Trends Genet.* **39**, 140–153 (2023).
20. Boller, S. & Grosschedl, R. The regulatory network of B-cell differentiation: a focused view of early B-cell factor 1 function. *Immunol. Rev.* **261**, 102–115 (2014).
21. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
22. Hystad, M. E. *et al.* Characterization of Early Stages of Human B Cell Development by Gene Expression Profiling. *J. Immunol.* **179**, 3662–3671 (2007).
23. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution. *Trends Immunol.* **43**, 210–229 (2022).
24. Mehtonen, J. *et al.* Single cell characterization of B-lymphoid differentiation and leukemic cell states during chemotherapy in ETV6-RUNX1-positive pediatric leukemia identifies drug-targetable transcription factor activities. *Genome Med. 2020 121* **12**, 1–25 (2020).
25. Semenchenko, K. *et al.* XRP44X, an Inhibitor of Ras/Erk Activation of the Transcription Factor Elk3, Inhibits Tumour Growth and Metastasis in Mice. *PLoS One* **11**, (2016).
26. Vijayakrishnan, J. *et al.* A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia* **31**, 573 (2017).
27. Zhang, J. *et al.* Methyltransferase-like protein 11A promotes migration of cervical cancer cells via up-regulating ELK3. *Pharmacol. Res.* **172**, (2021).
28. Lee, M., Cho, H. J., Park, K. S. & Jung, H. Y. ELK3 Controls Gastric Cancer Cell Migration and Invasion by Regulating ECM Remodeling-Related Genes. *Int. J. Mol. Sci.* **23**, (2022).
29. Liu, Z. *et al.* ELK3: A New Molecular Marker for the Diagnosis and Prognosis of Glioma. *Front. Oncol.* **11**, 1 (2021).
30. Yasuda, T. *et al.* Erk Kinases Link Pre-B Cell Receptor Signaling to Transcriptional Events Required for Early B Cell Expansion. *Immunity* **28**, 499–508 (2008).
31. Busslinger, M. Transcriptional Control of Early B Cell Development1. *https://doi.org/10.1146/annurev.immunol.22.012703.104807* **22**, 55–79 (2004).
32. Sipol, A. *et al.* MondoA drives malignancy in B-ALL through enhanced adaptation to metabolic stress. *Blood* **139**, 1184–1197 (2022).
33. Inaba, H. & Mullighan, C. G. Pediatric acute lymphoblastic leukemia. *Haematologica* **105**, 2524 (2020).
34. Li, J. F. *et al.* Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11711–E11720 (2018).
35. Wray, J. P. *et al.* Regulome analysis in B-acute lymphoblastic leukemia exposes Core Binding Factor addiction as a therapeutic vulnerability. *Nat. Commun. 2022 131* **13**, 1–18 (2022).
36. Vijayakrishnan, J. *et al.* Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. *Nat. Commun.* **9**, (2018).
37. Speedy, H. E. *et al.* Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nat. Commun.* **10**, (2019).
38. Liu, J. *et al.* The CEBPE rs2239633 genetic polymorphism on susceptibility to childhood acute lymphoblastic leukemia: an updated meta-analysis. *Environ. Health Prev. Med.* **26**, (2021).
39. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet. 2011 131* **13**, 59–69 (2011).
40. Xie, X. *et al.* Single-cell transcriptomic landscape of human blood cells. *Natl. Sci. Rev.* **8**,

2021 (2021).
41. Moroney, J. B., Vasudev, A., Pertsemlidis, A., Zan, H. & Casali, P. Integrative transcriptome and chromatin landscape analysis reveals distinct epigenetic regulations in human memory B cells. *Nat. Commun. 2020 111* **11**, 1–18 (2020).
42. Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6**, 377 (2010).
43. Singh, H. Transcriptional and epigenetic networks orchestrating immune cell development and function. *Immunol. Rev.* **261**, 5–8 (2014).
44. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
45. Fan, K., Pfister, E. & Weng, Z. Toward a comprehensive catalog of regulatory elements. *Hum. Genet. 2023* 1–21 (2023). doi:10.1007/S00439-023-02519-3
46. Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. ATAC2GRN: Optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC Genomics* **19**, 1–13 (2018).
47. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods 2017 1410* **14**, 975–978 (2017).
48. Ramírez, J., Lukin, K. & Hagman, J. From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment. *Curr. Opin. Immunol.* **22**, 177–184 (2010).
49. Liber, D. *et al.* Epigenetic Priming of a Pre-B Cell-Specific Enhancer through Binding of Sox2 and Foxd3 at the ESC Stage. *Cell Stem Cell* **7**, 114–126 (2010).
50. Chen, S. S. *et al.* Epigenetic alterations in a murine model for chronic lymphocytic leukemia. *Cell Cycle* **8**, 3663 (2009).
51. Mao, Y. *et al.* Silencing of ELK3 Induces S-M Phase Arrest and Apoptosis and Upregulates SERPINE1 Expression Reducing Migration in Prostate Cancer Cells. *Biomed Res. Int.* **2020**, (2020).
52. Blatti, C. *et al.* Bayesian machine learning enables identification of transcriptional network disruptions associated with drug-resistant prostate cancer. *Cancer Res.* **83**, 1361–1380 (2023).
53. Qin, Y. Z. *et al.* The Prognostic Significance of ZNF384 Fusions in Adult Ph-Negative B-Cell Precursor Acute Lymphoblastic Leukemia: A Comprehensive Cohort Study From a Single Chinese Center. *Front. Oncol.* **11**, 263 (2021).
54. Hirabayashi, S. *et al.* ZNF384-related fusion genes define a subgroup of childhood B-cell precursor acute lymphoblastic leukemia with a characteristic immunotype. *Haematologica* **102**, 118 (2017).
55. Jaitin, D. A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776 (2014).
56. Lavin, Y. *et al.* Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell* **169**, 750-765.e17 (2017).
57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
58. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139 (2010).
61. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289 (2016).
62. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.* **2**, 100141 (2021).
63. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods 2013 1012* **10**, 1213–1218 (2013).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

65. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Reports 2019 91* **9**, 1–5 (2019).
66. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 1–9 (2008).
67. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
68. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
69. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187 (2014).
70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
71. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
72. Wieland, T. REAT: A Regional Economic Analysis Toolbox for R. *REGION* **7**, R1–R57 (2019).
73. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* **11**, (2012).
74. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 1–21 (2019).
75. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gku727
76. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718 (2011).
77. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
78. Csardi, G. & Tamas Nepusz. The igraph software package for complex network research. *InterJournal, complex Syst.* **1695**, 1–9 (2006).
79. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
80. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, (2017).
81. Vidal, E. *et al.* OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.* **46**, e49 (2018).
82. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, (2016).
83. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat. Protoc.* **15**, 991–1012 (2020).
84. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
85. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: An R Package for Analyzing Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2010).
86. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, (2020).
87. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882 (2012).
88. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
89. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol. 2015 335* **33**, 495–502 (2015).
90. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol. 2018 365* **36**, 411–420 (2018).
91. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
92. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

93. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol. 2019 202* **20**, 163–172 (2019).
94. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods 2021 1811* **18**, 1333–1341 (2021).
95. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
96. Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single cell chromatin accessibility data. *Mol. Cell* **71**, 858 (2018).

Figure legends

Figure 1. Overview of chromatin and RNA profiling and cis-Regulatory Elements identification.

(A) Schematic representation of data generated in the project.
(B) Principal Component Analysis of RNA-Seq and ATAC-Seq data. Each dot represents a sample. Color-code by cell subpopulation.
(C) Representation of ATAC-Seq signal for PAX5 genomic region for the 8 cell subpopulations, all at the same scale. Bar plots on the right side show mRNA levels. Histone marks from ENOCDE data are shown in the bottom part. Consensus OCRs are shown in the top part in black.
(D) Dot plot representation of gene expression and chromatin accessibility for PAX5 gene and their promoter region.
(E) Heatmap representation of chromatin accessibility for cCREs identified for PAX5 gene (n=46). Bar plot representation of the PAX5 expression is shown at the top of the heatmap. Distance from TSS, rho and p-value of Spearman correlation between OCRs vs PAX5 gene, and the overlap with histone marks and PC-HiC interactions are shown at the right side.
(F) The same as (D) but for the PAX5 gene and a previously reported enhancer and new candidate enhancer elements (G).
(G) The same as (D) but for previously reported enhancer and new candidate enhancer elements of PAX5.
(H) Chromatin accessibility distribution of cCREs of PAX5 in Naïve B cells (CD5+ and CD5-) split by OCRs previously reported (light grey) or newly identified (grey).

Figure 2. Gene Regulatory Network of early B cell differentiation.

(A) Heatmap representation of genes included in the GRN (n=7,310 genes). On the left is the Z-score of gene expression values. On the right is a binary representation of gene expression patterns, with high expression in brown.
(B) Heatmap representation of OCRs included in the GRN (n=16,074 OCRs). On the left is the Z-score of chromatin accessibility values. On the right is a binary representation of the chromatin accessibility patterns, with high accessibility in brown.
(C) TFBS activity score shown as Z-score for all TFs within the GRN (n=169 TFs).
(D) Multidimensional Scaling representation from the Pearson correlation of cell subpopulation specific GRNs. Color-code by cell subpopulation.
(E) Degree and Betweenness GRNs properties representation for the whole and cell subpopulation specific GRNs. TFs with the highest degree are shown in yellow, and those with the highest betweenness in cyan.
(F) Zoom-in from (E) panel showing the top 10 TF for each cell subpopulation.
(G) Density plots of TFBS activity score (left) and gene expression (right) for ELK3 for each cell subpopulation.

Figure 3. Upstream regulation dynamics of specific genes along B cell differentiation.

(A) Network representation of upstream regulators identified for EBF1 gene at CLP, pro-B and pre-B stage. Node size and color show the gene expression signal and the edge color and width show the TF activity. Activators are shown in red and repressors in blue,

and as strong the correlation wider the edge. Median values per cell subpopulation are represented.

(B) The same as (A) but for the ELK3 gene.
(C) Heatmap representation of gene expression for TF regulators of EBF1. Bar plot representation of the EBF1 expression is shown at the top of the heatmap. Rho values of Spearman correlation for each TF vs EBF1 are shown on the right side.
(D) The same as (C) but for the ELK3 gene.
(E) Density plots of TFBS activity score (top) and gene expression (bottom) for MLXIP for each cell subpopulation.

Figure 4. Uncovering ALL through epigenetic and regulatory data.

(A) Schematic representation of ALL public data explored.
(B) Over Representation Analysis (ORA) of BCP ALL subtypes biomarkers over the biomarkers of different B cell subpopulations. Dot size represents the –log10(adjusted p-value) and the color the gene ratio.
(C) ORA for BCP ALL subtypes G1 (MEF2D fusion), G2 (TCF3-PBX1), and G3 (ETV6-RUNX1) over the set of genes regulated by MEFD2, TCF3 and PBX1 or ETV6 and RUNX1, respectively, within each cell subpopulation. Dot size represents the –log10(adjusted p-value) and the color the gene ratio.
(D) ORA for RUNX1-ETV6 high affinity binding sites (n=711) over the consensus OCRs for each cell subpopulation. Dot size represents the –log10(adjusted p-value), and the color the odds ratio.
(E) Violin plot of ELK3 gene expression by BCP ALL subtypes. Color-code by BCP ALL subtypes.
(F) Representation of ATAC-Seq signal for rs17481869 genomic region for the 8 cell subpopulations, all at the same scale. Histone marks from ENOCDE data are shown in the bottom part. Consensus OCRs are shown in the top part in black.
(G) The same as (F) but for the loci within the ELK3 gene (rs2268508, rs17736737, and rs4762284).
(H) The same as (F) but for the rs2239633 locus, located in the CEBPE promoter region.
(I) The same as (E) but for the CEBPE gene.
(J) The same as (F) but for the rs35837782 locus.
(K) Density plot of gene expression by cell subpopulation for LHPP, FAM53B-AS1 and CHST15 gene.
(L) The same as (E) but for LHPP and CHST15 gene.

Figure 5. Chromatin accessibility and mRNA coordination.

(A) Dot plot showing the promoter activation. For each induced gene over B cell differentiation, the dots denote the cell transition at which the gene mRNA shows the highest increase versus the cell transition at which the gene promoter region shows the highest accessibility (left plot). The same is shown but for the genes that are block over B cell differentiation (right plot).
(B) The same as (A) but showing the enhancer activation.
(C) UMAP representation of single-cell RNA-Seq (left) and single-cell ATAC-Seq (right) data where each dot is a cell. On the top panel, cells are colored according to cell subpopulation annotation; in the bottom, cells are colored according to pseudotime score and pseudotime trajectory is shown as a black line.
(D) Dot plot showing the expression of the PAX5 gene over time. Each dot shows the proportion of cells expressing the gene at a time point. The proportion of cells without

signal is shown in grey. The bottom dot plot shows the same but for the chromatin accessibility of the PAX5 promoter. As a dot collapse more than one cell, the mean gene expression or chromatin accessibility is represented. Dots with gene or OCR signal are colored by the majoritarian cell subpopulation.

(E) Representation of ATAC-Seq signal for CFH promoter region for the 8 cell subpopulations, all at the same scale. Bar plots on the right side show mRNA levels. Histone marks from ENOCDE data are shown in the bottom part. Consensus OCRs are shown in the top part in black.

(F) The same as (D) but for the CFH gene and its promoter region.

(G) The same as (E) but for the cCRE of the CDH7 gene.

(H) The same as (D) but for the CDH7 gene and its candidate enhancer region.

(I) UMAP representation of single-cell RNA-Seq data where each dot is a cell. Cells are colored according to signature score. Left plot show the enrichment of HSC and CLP signature, and right plot of pro-B marker genes.

(J) UMAP representation of single-cell ATAC-Seq data where each dot is a cell. Cells are colored according to signature score. Left plot show the enrichment derived from the promoters of HSC and CLP marker genes, and right plot from the promoters of pro-B marker genes.

(K) The same as (J) but instead to show the enrichment of promoters shows the enrichment based on enhancer OCRs.

Supplementary Figures

Figure S1. Gating strategy

(A) Gating strategy followed to obtain the different B cell subpopulations from bone marrow samples.
(B) Gating strategy followed to obtain the single-cell sample from the bone marrow.

Figure S2. Identification of cis-regulatory elements in early B-cell differentiation.

(A) Distribution of OCR genetic annotation. Each color represents a genomic region and the bar height the absolute frequency of each annotation.
(B) Schematic representation of strategy followed for cis-regulatory elements identification.
(C) Veen diagram of the cis-regulation identified for each gene. The total number of expressed genes without cis-regulatory elements identified are shown in yellow, the expressed genes with only promoter elements identified are shown in green, expressed genes with only enhancer elements identified are shown in purple and the genes with both, promoter and enhancer elements identified are shown in blue.
(D) Barplot representation of the number of promoter OCRs significantly associated to at gene.
(E) Barplot representation of the number of distal OCRs (intronic or distal intergenic) significantly associated to a gene.
(F) Distribution of the distance to TSS for the enhancer CREs by different Spearman correlation rho values.
(G) Barplot representation of significant cis-regulatory elements (cCREs) overlapping histone marks, shown as proportions. Proportion of OCRs without histone marks information are shown in pink (no histone mark). Barplots are shown for all OCRs, those defined as cCREs and those not defined as cCREs.
(H) Barplot representation of significant cis-regulatory elements (cCREs) overlapping Hi-C compartments; shown as proportions. Hi-C compartments are differentiated between active compartment (A, in coral), intermediate compartment (I, in blue) and inactive compartment (B, in green). Proportion of OCRs without compartment information are shown in purple (missing). Barplots are shown for all OCRs, those defined as cCREs and those not defined as cCREs.
(I) Distribution of chromatin accessibility for cCREs in Naïve B cells split by new CREs and CREs previously described in PC-HiC data.
(J) Heatmap representation of 2,161 OCR-gene pairs overlapping with PC-HiC public interactions. Left heatmap shows the chromatin accessibility levels (in z-score) and right heatmap the gene expression level. Each row corresponds and OCR-gene pair. Color-coding according cell subpopulation is shown on the top row of the heatmap. High accessible regions are shown in red and low accessible region in black. Up-regulated genes are shown in red and down-regulated genes in blue. Finally, some OCR-gene pairs are highlighted on the right.

Figure S3. Transcriptional and epigenetic characterization of early B-cell differentiation.

(A) Heatmap representation of Pearson correlation matrices (rho) between samples based on RNA-seq data (top left), ATAC-seq data (top right), ATAC-seq data at all promoter OCRs (bottom left), ATAC-seq data at all intronic OCRs (bottom center) and ATAC-seq data at all distal intergenic OCRs (bottom right). The corresponding OCR proportion of

each OCR-type (promoter, intronic, distal intergenic) is shown. Color-coding according cell subpopulation is shown on the top row of each heatmap.

(B) A t-SNE representation of OCRs identified in the study. First, Gini index for all OCRs; Gini index highlights OCR variability. OCRs with large variability are shown in blue and those with low variability in red. Following OCR distribution for promoter OCRs, intronic OCRs and distal intergenic OCRs. High density regions are shown in yellow and low density in dark blue.

(C) Variance component decomposition of the mRNA expression for every gene (as column), in a variance component model that discretizes the explanatory power of promoter OCRs (in green), distal enhancer OCRs (in blue) and unexplained variance (in red). The proportion of variance explained has been shown for the total number of identified genes (top left), for differentially expressed genes (bottom left), and for the genes not differentially expressed (bottom right). Also, results of the variance components model for two permuted matrices per gene model are shown (top right).

(D) Heatmap representation of significant differential expressed genes according cell subpopulation transitions (11,080 genes). Significant genes were defined as a |logFC|≥1.5 and adjusted p-value≤0.05. Z-score of normalized gene expression levels is represented. Up-regulated genes are shown in red and down-regulated genes in blue. Each row shows one individual gene and each column an individual sample. Color-coding according cell subpopulation is shown on the top row of the heatmap. Genes are clustered according to the different gene expression pattern. Additional annotation on the right side of the heatmap highlight the cell subpopulations that show over-expression for the different clusters (brown cells). Finally, key genes are highlighted on the right.

(E) Representation of top ten enriched GO terms over B cell transitions. Normalized enrichment score is shown in a red-blue color range for each term and the –log10 of adjusted p-value as dot size.

(F) Heatmap representation of significant differential accessible regions according cell subpopulation transitions (77,559 OCRs). Significant OCRs were defined as an adjusted p-value≤0.01. Z-score of normalized chromatin accessibility levels is represented. High accessible regions are shown in green and low accessible region in purple. Each row shows one individual OCR and each column an individual sample. Color-coding according cell subpopulation is shown on the top row of the heatmap. OCRs are clustered according to the different chromatin accessibility patterns. Additional annotation on the right side of the heatmap highlight the cell subpopulations that show high accessibility for the different clusters (brown cells).

(G) Heatmap representation of TFBS accessibility score (left) and gene expression level (right) of TFs identified within cCREs (n=169). Z-score of both measures are shown. Highly accessible scores are shown in orange and low accessible scores in blue. Up-regulated genes are shown in red and down-regulated genes in blue. Each row shows one individual TF and each column an individual sample. Color-coding according cell subpopulations is shown on the top row of the heatmap. Additional row annotation is shown between heatmaps, showing the TF motif variability, the coefficient of variation of gene expression and the Spearman's correlation coefficient. Finally, on the right side of the gene expression heatmap some key TFs on B cell differentiation are highlighted.

Figure S4. Characterization of ELK3 and FOXD3.

(A) Representation of the enriched GO terms derived from the genes regulated by ELK3 at pro-B and pre-B state. Adjusted p-value (in –log10) is shown in a red-blue color range for each term and the number of genes overlapping each term as dot size. Gene ratio is shown in the x-axis.

(B) Representation of the enrichment plot for the GO term GO:0016477 (cell migration) enriched within genes regulated by ELK3 at pro-B and pre-B state (fold-change between these stages is provided to compute the enrichment).

(C) TFBS variability for each TFs.

(D) Density plots of gene expression (top) and TFBS activity score (bottom) for FOXD3 for each cell subpopulation.
(E) Representation of ATAC-Seq signal for FOXD3 genomic region for the 8 cell subpopulations, all at the same scale. Bar plots on the right side show mRNA levels. Histone marks from ENOCDE data are shown in the bottom part. Consensus OCRs are shown in the top part in black. Highlighted the OCR region corresponding to FOXD3 promoter.
(F) Representation of the enrichment plot for the GO term GO:0044772 (mitotic cell cycle phase transition) enriched within genes regulated by FOXD3 at pre-B and Immature B state (fold-change between these stages is provided to compute the enrichment).

Figure S5. Gene Regulatory Network of early B cell differentiation.

(A) Multidimensional Scaling representation from the Pearson correlation of cell subpopulation specific TF-TF GRNs. Color-code by cell subpopulation.
(B) Distribution of the number of upstream regulators for each gene (left) and for the downstream regulated genes for each TF (left).
(C) Network representation of upstream regulators identified for EBF1 gene for each cell subpopulation. Node size and color show the gene expression signal and the edge color and width show the TF activity. Activators are shown in red and repressors in blue, and as strong the correlation wider the edge. Median values per cell subpopulation are represented.
(D) The same as (C) but for the ELK3 gene.

Figure S6. Gene Regulatory Network of early B cell differentiation over transitions.

Heatmaps representation of transitions between TF-TF gene regulatory networks based on adjacent matrices. Representation for 6 transitions are shown; HSC to CLP, CLP to proB, proB to preB, preB to Immature B, Immature B to Transitional B, and Transitional B to Naïve B cells (CD5+). Regulator TFs are shown by rows and target TFs by columns. TF-TF interactions present in both GRNs from target transitions are shown in light brown, TF-TF interactions present only in the earlier stage of the transition are shown in dark green and TF-TF interactions present only in the later stage of the transition are shown in dark brown. Mean gene expression levels for each TF and cell subpopulation are shown on the rows of the heatmap and at the right side. High expressed genes are shown in red and low expressed genes in blue. Also, TF gene symbols are shown.

Figure S7. Chromatin accessibility and mRNA coordination.

(A) Distribution of correlation coefficient and adjusted p-value (shown in –log10) for the Spearman correlation output of gene expression and promoter accessibility.
(B) Dot plot showing the enhancer activation for interactions previously described under PC-HiC approach in Naïve B cells. For each induced gene over B cell differentiation, the dots denote the cell transition at which the gene mRNA shows the highest increase versus the cell transition at which the gene enhancer region shows the highest accessibility (left plot). The same is shown but for the genes that are block over B cell differentiation (right plot).

Figure S8. Single-cell data of early B-cell differentiation

(A) UMAP plot from scRNA-seq data showing the distribution of the single cells and its annotation. Three cell subpopulations are identified; HSCs is green, CLPs in turquoise and proB cells in dark blue.

(B) UMAP plot from scATAC-seq data showing the distribution of the single cells and it's annotation. Three cell subpopulations are identified; HSCs is green, CLPs in turquoise and proB cells in dark blue. On the left UMAP of scATAC-seq based on counts from the OCRs identified within single-cell data, on the right UMAP of scATAC-seq based on counts from the OCRs identified in bulk data.

(C) UMAP plots from scRNA-seq data colored according to gene expression signature of marker genes per each cell subpopulation (according to expression patterns shown in Figure S3D). Cells with increased signature levels are shown in blue.

(D) UMAP plots from scATAC-seq data colored according to chromatin accessibility signature of marker OCRs per each cell subpopulation (according to accessibility patterns shown in Figure S3F). Cells with increased signature levels are shown in green.

(E) B-cell markers gene expression. The size and color of each circle represents the average gene expression of each cell subpopulation and the percentage of cells within each cell subpopulation with gene expression.

Figure 1

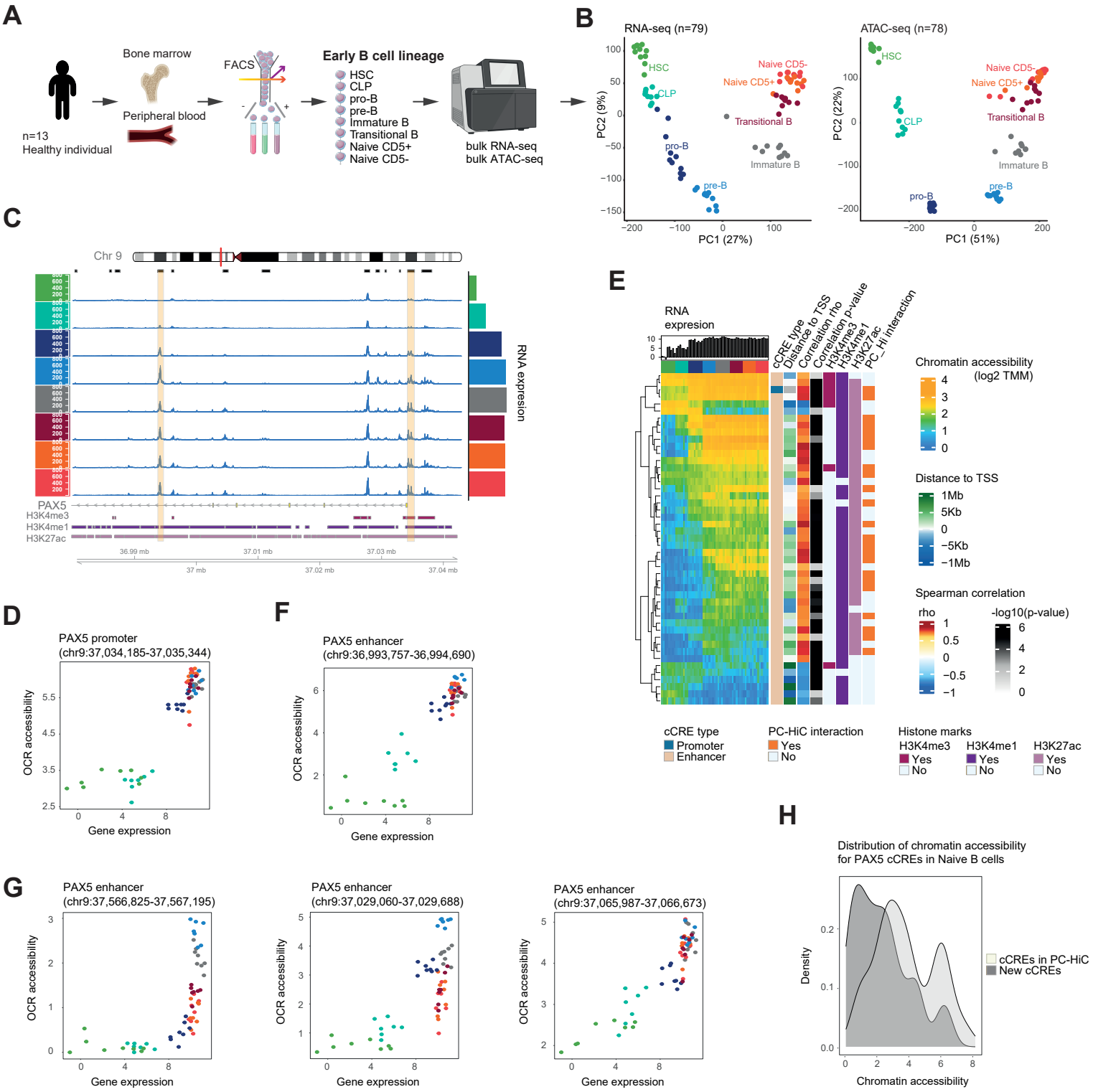

A
Bone marrow
Peripheral blood
n=13
Healthy individual
FACS
Early B cell lineage
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5+
Naive CD5-
bulk RNA-seq
bulk ATAC-seq
B
RNA-seq (n=79)
ATAC-seq (n=78)
PC1 (27%)
PC2 (9%)
PC1 (51%)
PC2 (22%)
C
Chr 9
RNA expresion
PAX5
H3K4me3
H3K4me1
H3K27ac
36.99 mb
37.01 mb
37.03 mb
37 mb
37.02 mb
37.04 mb
D
PAX5 promoter
(chr9:37,034,185-37,035,344)
OCR accessibility
Gene expression
E
RNA
expresion
cCRE type
Distance to TSS
Correlation rho
Correlation p-value
H3K4me3
H3K4me1
H3K27ac
PC_HI interaction
Chromatin accessibility
(log2 TMM)
Distance to TSS
1Mb
5Kb
0
−5Kb
−1Mb
Spearman correlation
rho
-log10(p-value)
cCRE type
Promoter
Enhancer
PC-HiC interaction
Yes
No
Histone marks
H3K4me3
H3K4me1
H3K27ac
F
PAX5 enhancer
(chr9:36,993,757-36,994,690)
OCR accessibility
Gene expression
G
PAX5 enhancer
(chr9:37,566,825-37,567,195)
PAX5 enhancer
(chr9:37,029,060-37,029,688)
PAX5 enhancer
(chr9:37,065,987-37,066,673)
OCR accessibility
Gene expression
H
Distribution of chromatin accessibility
for PAX5 cCREs in Naive B cells
Density
Chromatin accessibility
cCREs in PC-HiC
New cCREs

Figure 2

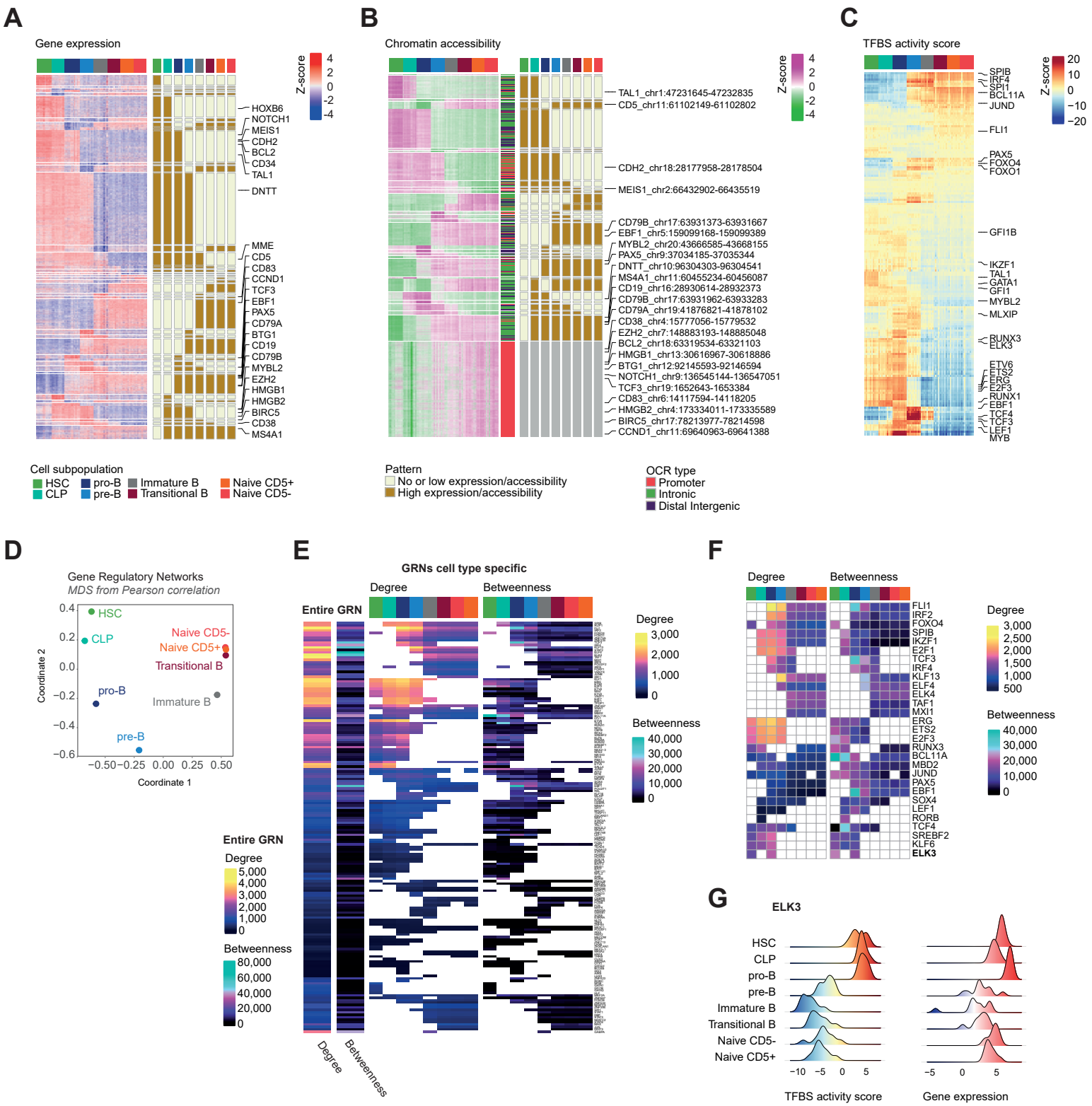

A
Gene expression
Z-score
4
2
0
-2
-4
HOXB6
NOTCH1
MEIS1
CDH2
BCL2
CD34
TAL1
DNTT
MME
CD5
CD83
CCND1
TCF3
EBF1
PAX5
CD79A
BTG1
CD19
CD79B
MYBL2
EZH2
HMGB1
HMGB2
BIRC5
CD38
MS4A1
Cell subpopulation
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5+
Naive CD5-
B
Chromatin accessibility
Z-score
4
2
0
-2
-4
TAL1_chr1:47231645-47232835
CD5_chr11:61102149-61102802
CDH2_chr18:28177958-28178504
MEIS1_chr2:66432902-66435519
CD79B_chr17:63931373-63931667
EBF1_chr5:159099168-159099389
MYBL2_chr20:43666585-43668155
PAX5_chr9:37034185-37035344
DNTT_chr10:96304303-96304541
MS4A1_chr11:60455234-60456087
CD19_chr16:28930614-28932373
CD79B_chr17:63931962-63933283
CD79A_chr19:41876821-41878102
CD38_chr4:15777056-15779532
EZH2_chr7:148883193-148885048
BCL2_chr18:63319534-63321103
HMGB1_chr13:30616967-30618886
BTG1_chr12:92145593-92146594
NOTCH1_chr9:136545144-136547051
TCF3_chr19:1652643-1653384
CD83_chr6:14117594-14118205
HMGB2_chr4:173334011-173335589
BIRC5_chr17:78213977-78214598
CCND1_chr11:69640963-69641388
Pattern
No or low expression/accessibility
High expression/accessibility
OCR type
Promoter
Intronic
Distal Intergenic
C
TFBS activity score
Z-score
20
10
0
−10
−20
SPIB
IRF4
SPI1
BCL11A
JUND
FLI1
PAX5
FOXO4
FOXO1
GFI1B
IKZF1
TAL1
GATA1
GFI1
MYBL2
MLXIP
BUNX3
ELK3
ETV6
ETS2
ERG
E2F3
RUNX1
EBF1
TCF4
TCF3
LEF1
MYB
D
Gene Regulatory Networks
MDS from Pearson correlation
HSC
CLP
Naive CD5-
Naive CD5+
Transitional B
pro-B
Immature B
pre-B
Coordinate 2
Coordinate 1
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.50
−0.25
0.00
0.25
0.50
E
GRNs cell type specific
Degree
Betweenness
Entire GRN
Degree
3,000
2,000
1,000
0
Betweenness
40,000
30,000
20,000
10,000
0
Entire GRN
Degree
5,000
4,000
3,000
2,000
1,000
0
Betweenness
80,000
60,000
40,000
20,000
0
Degree
Betweenness
F
Degree
Betweenness
FLI1
IRF2
FOXO4
SPIB
IKZF1
E2F1
TCF3
IRF4
KLF13
ELF4
ELK4
TAF1
MXI1
ERG
ETS2
E2F3
RUNX3
BCL11A
MBD2
JUND
PAX5
EBF1
SOX4
LEF1
RORB
TCF4
SREBF2
KLF6
ELK3
Degree
3,000
2,500
2,000
1,500
1,000
500
Betweenness
40,000
30,000
20,000
10,000
0
G
ELK3
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5-
Naive CD5+
−10
−5
0
5
TFBS activity score
−5
0
5
Gene expression

Figure 3

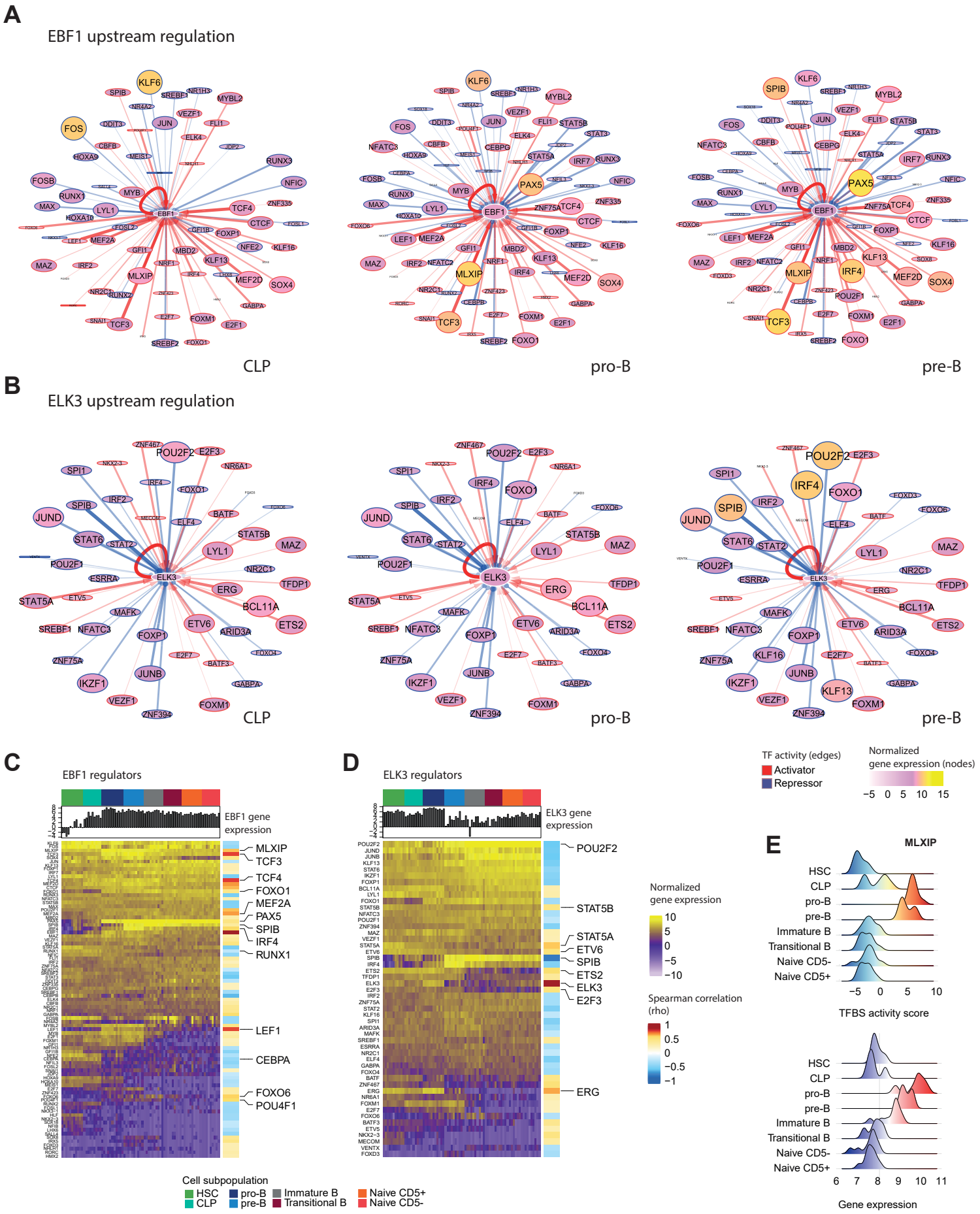

A
EBF1 upstream regulation
CLP
pro-B
pre-B
B
ELK3 upstream regulation
CLP
pro-B
pre-B
C
EBF1 regulators
EBF1 gene expression
MLXIP
TCF3
TCF4
FOXO1
MEF2A
PAX5
SPIB
IRF4
RUNX1
LEF1
CEBPA
FOXO6
POU4F1
D
ELK3 regulators
ELK3 gene expression
POU2F2
STAT5B
STAT5A
ETV6
SPIB
ETS2
ELK3
E2F3
ERG
Normalized gene expression
Spearman correlation (rho)
TF activity (edges)
Activator
Repressor
Normalized gene expression (nodes)
E
MLXIP
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5-
Naive CD5+
TFBS activity score
Gene expression
Cell subpopulation
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5+
Naive CD5-

Figure 4

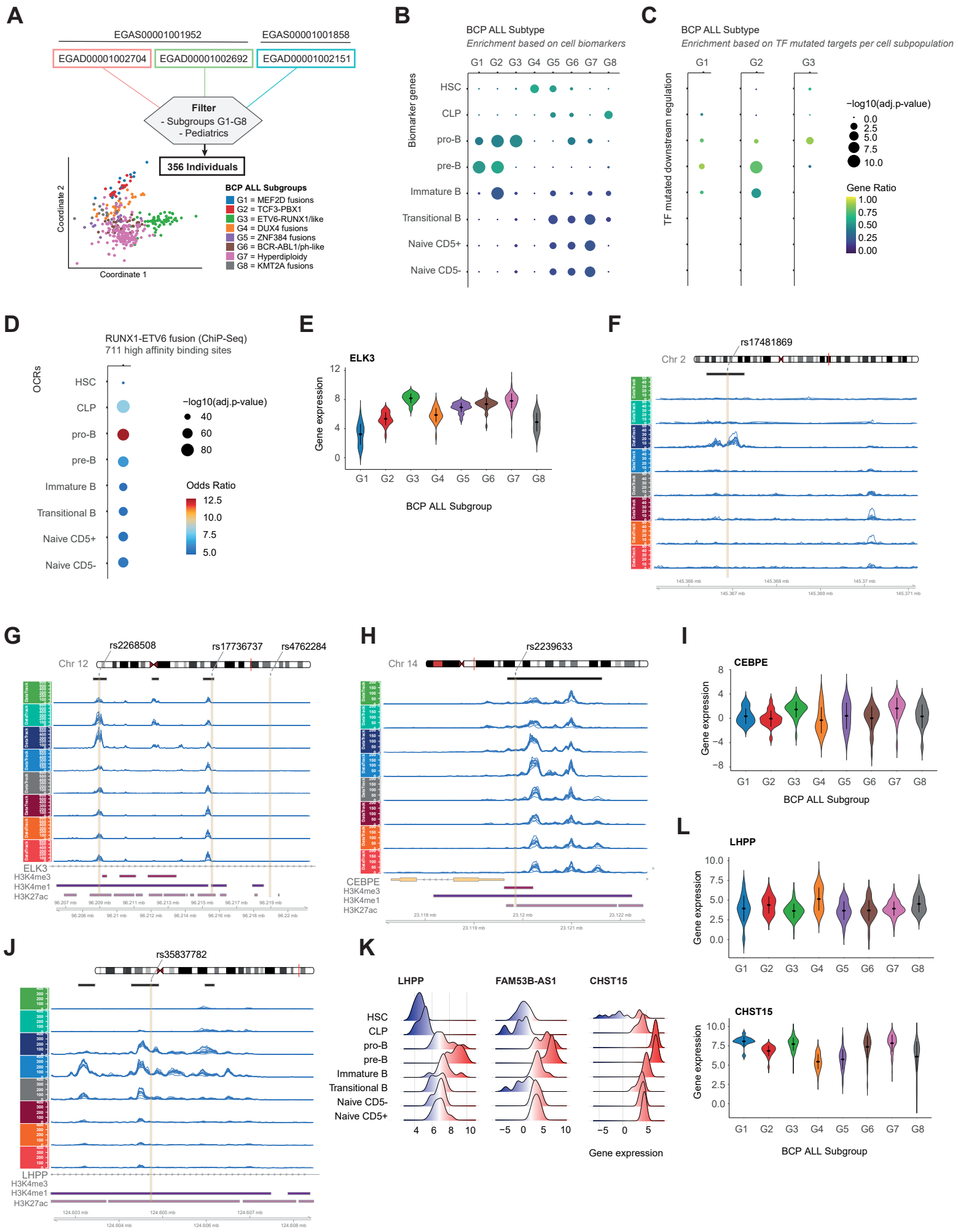
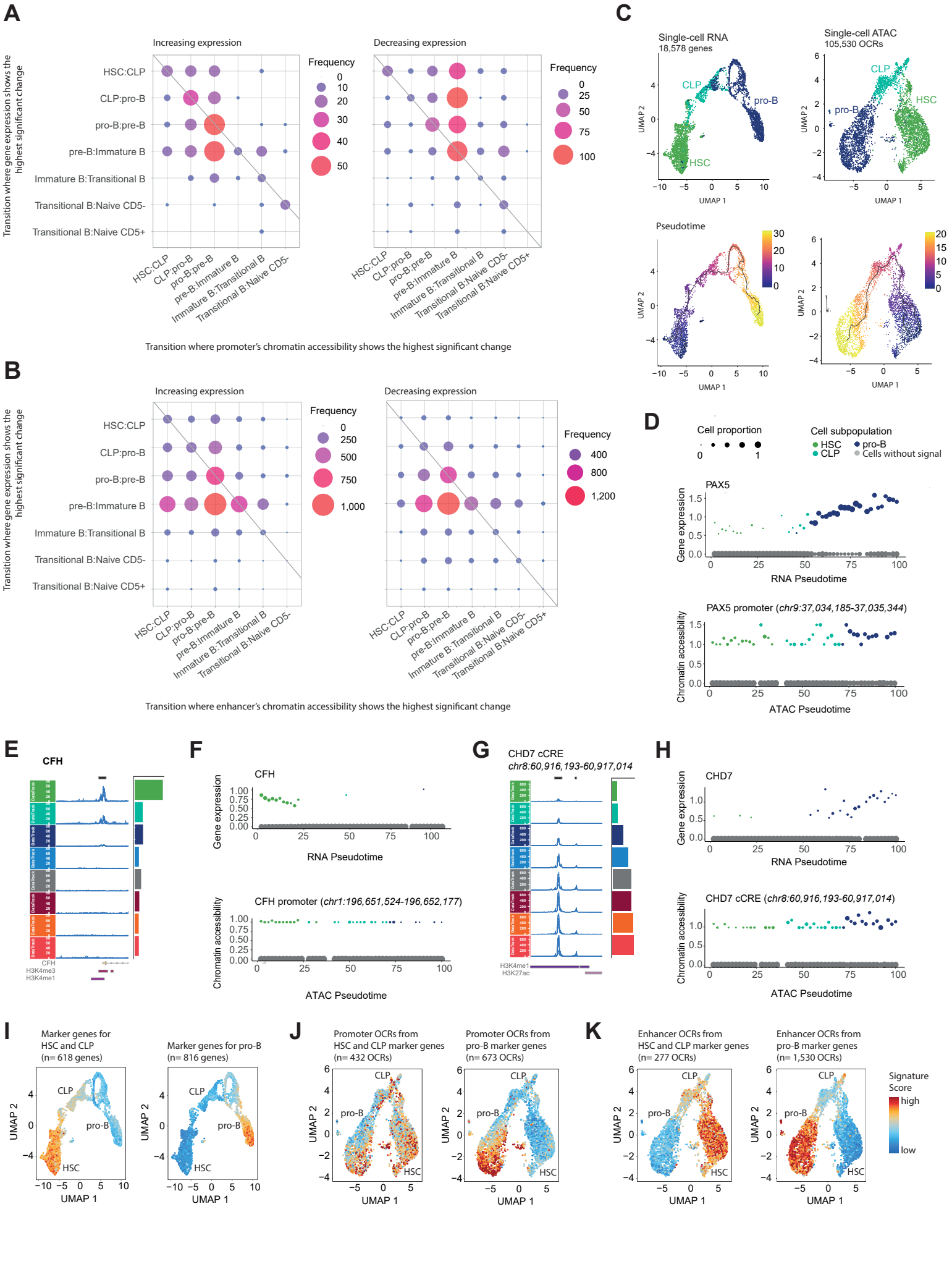

A
EGAS00001001952
EGAS00001001858
EGAD00001002704
EGAD00001002692
EGAD00001002151
Filter
- Subgroups G1-G8
- Pediatrics
356 Individuals
Coordinate 2
Coordinate 1
BCP ALL Subgroups
G1 = MEF2D fusions
G2 = TCF3-PBX1
G3 = ETV6-RUNX1/like
G4 = DUX4 fusions
G5 = ZNF384 fusions
G6 = BCR-ABL1/ph-like
G7 = Hyperdiploidy
G8 = KMT2A fusions
B
BCP ALL Subtype
Enrichment based on cell biomarkers
Biomarker genes
HSC
CLP
pro-B
pre-B
Immature B
Transitional B
Naive CD5+
Naive CD5-
C
BCP ALL Subtype
Enrichment based on TF mutated targets per cell subpopulation
TF mutated downstream regulation
−log10(adj.p-value)
Gene Ratio
D
RUNX1-ETV6 fusion (ChiP-Seq)
711 high affinity binding sites
OCRs
−log10(adj.p-value)
Odds Ratio
E
ELK3
Gene expression
BCP ALL Subgroup
F
rs17481869
Chr 2
G
rs2268508
rs17736737
rs4762284
Chr 12
ELK3
H3K4me3
H3K4me1
H3K27ac
H
rs2239633
Chr 14
CEBPE
I
CEBPE
L
LHPP
CHST15
J
rs35837782
LHPP
K
LHPP
FAM53B-AS1
CHST15
Gene expression

Figure 5


A
Increasing expression
Decreasing expression
Frequency
Transition where gene expression shows the highest significant change
HSC:CLP
CLP:pro-B
pro-B:pre-B
pre-B:Immature B
Immature B:Transitional B
Transitional B:Naive CD5-
Transitional B:Naive CD5+
Transition where promoter's chromatin accessibility shows the highest significant change
B
Transition where enhancer's chromatin accessibility shows the highest significant change
C
Single-cell RNA
18,578 genes
Single-cell ATAC
105,530 OCRs
Pseudotime
UMAP 1
UMAP 2
D
Cell proportion
Cell subpopulation
HSC
pro-B
CLP
Cells without signal
PAX5
Gene expression
RNA Pseudotime
PAX5 promoter (chr9:37,034,185-37,035,344)
Chromatin accessibility
ATAC Pseudotime
E
CFH
F
CFH promoter (chr1:196,651,524-196,652,177)
G
CHD7 cCRE
chr8:60,916,193-60,917,014
H
CHD7
CHD7 cCRE (chr8:60,916,193-60,917,014)
I
Marker genes for HSC and CLP (n= 618 genes)
Marker genes for pro-B (n= 816 genes)
J
Promoter OCRs from HSC and CLP marker genes (n= 432 OCRs)
Promoter OCRs from pro-B marker genes (n= 673 OCRs)
K
Enhancer OCRs from HSC and CLP marker genes (n= 277 OCRs)
Enhancer OCRs from pro-B marker genes (n= 1,530 OCRs)
Signature Score
high
low